Final Project

Raymundo Lopez

2024-12-08

Data Set #1 - Pricing Cars

For this exercise I'm going to go ahead and fit a logistic regression model mapping price as a function of the other variables in this data set.

As far as data transformations go, we will be doing the following:

- Drop Column 1 (or just omit it from our training)
- Convert trim to a factor datatype so our model knows it's categorical.
- Drop subTrim as it's mostly missing and only describes hybrid or not, but that's specififed in the fuel type already.
- Convert condition to a factor datatype.
- Convert isOneOwner to 1s and 0s
- Convert color to factor datatype.
- Convert displacement to numerical data.
- Convert fuel, state, region, soundSystem, and wheelType to factor datatype.
- Drop wheelSize as there are too many missing types.
- featureCount and price will remain unchanged.

```
## # A tibble: 10 x 17
##     ...1 trim  subTrim condition isOneOwner mileage  year color  displacement
##    <dbl> <fct> <fct>   <fct>          <dbl>   <dbl> <dbl> <fct>         <dbl>
##  1     2 320   unsp    Used               0  129948  1995 Gold            3.2
##  2     4 320   unsp    Used               0  140428  1997 White           3.2
##  3     7 420   unsp    Used               0  113622  1999 Silver          4.2
##  4     8 420   unsp    Used               0  167673  1999 Silver          4.2
##  5    11 500   unsp    Used               0   63457  1997 Silver          5
##  6    13 430   unsp    Used               0   82419  2002 White           4.3
##  7    17 430   unsp    Used               0  101264  2000 White           4.3
##  8    18 430   unsp    Used               0  110651  2001 Black           4.3
##  9    19 430   unsp    Used               0  108173  2000 Silver          4.3
## 10    21 430   unsp    Used               0  187977  2002 Gray            4.3
## # i 8 more variables: fuel <fct>, state <fct>, region <fct>, soundSystem <fct>,
## #   wheelType <fct>, wheelSize <chr>, featureCount <dbl>, price <dbl>
```

Due to the size of our data set and prominence of categorical data, I don't think any visual exploratory work such as a scatter plot matrix will be very useful, and since so much is non-numerical we can't do a correlation matrix either. Instead, I'm going to jump straight into training our MLR.

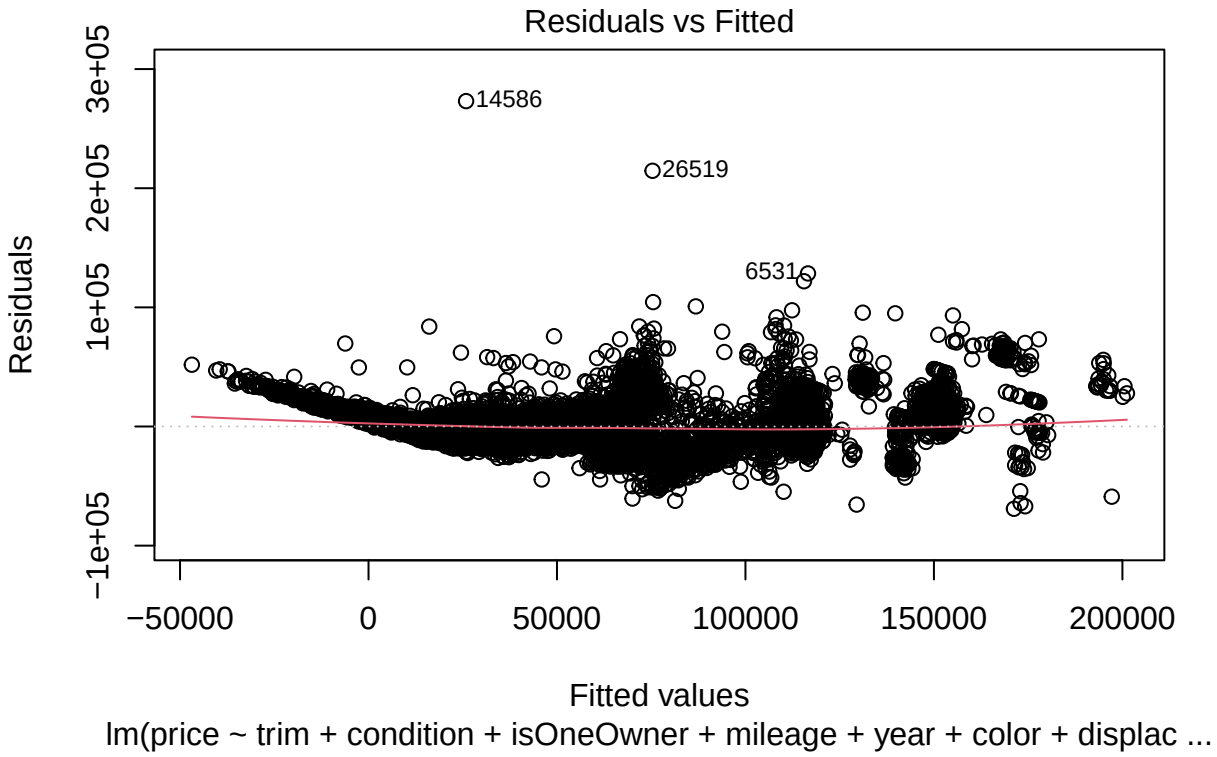
```
##
## Call:
## lm(formula = price ~ trim + condition + isOneOwner + mileage +
##     year + color + displacement + fuel + state + region + soundSystem +
##     wheelType + featureCount, data = carsData)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -69063  -5678   -958   3778 273115
##
## Coefficients: (9 not defined because of singularities)
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -8.821e+06  1.036e+05 -85.171  < 2e-16 ***
## trim350                    -3.461e+04  1.472e+03 -23.507  < 2e-16 ***
## trim400                    -2.821e+04  1.161e+04  -2.429 0.015152 *
## trim420                     4.125e+03  1.244e+03   3.315 0.000918 ***
## trim430                    -2.086e+04  1.012e+03 -20.605  < 2e-16 ***
## trim450                    -3.476e+04  1.341e+04  -2.591 0.009570 **
## trim500                    -1.613e+04  1.114e+03 -14.479  < 2e-16 ***
## trim55 AMG                 -1.618e+04  1.389e+03 -11.655  < 2e-16 ***
## trim550                    -2.473e+04  1.344e+03 -18.401  < 2e-16 ***
## trim600                    -5.759e+03  1.428e+03  -4.033 5.52e-05 ***
## trim63 AMG                  1.370e+04  1.561e+03   8.778  < 2e-16 ***
## trim65 AMG                  3.140e+04  1.671e+03  18.790  < 2e-16 ***
## trimunsp                    3.286e+04  2.373e+03  13.850  < 2e-16 ***
## conditionNew                3.650e+04  2.806e+02 130.051  < 2e-16 ***
## conditionUsed              -4.990e+03  2.490e+02 -20.042  < 2e-16 ***
## isOneOwner                 -1.986e+02  2.144e+02  -0.926 0.354194
## mileage                    -1.334e-01  3.150e-03 -42.346  < 2e-16 ***
## year                        4.437e+03  5.122e+01  86.631  < 2e-16 ***
## colorBlack                 -3.898e+02  8.086e+02  -0.482 0.629715
## colorBlue                  -1.025e+03  8.523e+02  -1.203 0.228944
## colorBronze                 4.596e+03  4.171e+03   1.102 0.270555
## colorBrown                  2.281e+02  1.649e+03   0.138 0.889953
## colorGold                   1.039e+03  1.088e+03   0.955 0.339813
## colorGray                  -1.434e+03  8.425e+02  -1.703 0.088669 .
## colorGreen                  4.408e+01  1.194e+03   0.037 0.970547
## colorPurple                 6.553e+03  4.178e+03   1.568 0.116815
## colorRed                   -1.738e+02  1.030e+03  -0.169 0.866053
## colorSilver                -1.293e+03  8.138e+02  -1.588 0.112227
## colorTurquoise             -5.424e+02  5.237e+03  -0.104 0.917506
## colorunsp                   5.974e+02  8.587e+02   0.696 0.486618
## colorWhite                  1.301e+03  8.215e+02   1.584 0.113202
## colorYellow                -6.102e+03  8.217e+03  -0.743 0.457752
## displacement               -2.787e+03  2.975e+02  -9.369  < 2e-16 ***
## fuelGasoline               -2.551e+03  1.305e+03  -1.954 0.050723 .
## fuelHybrid                 -9.870e+03  1.167e+04  -0.846 0.397761
## fuelunsp                    2.309e+04  1.583e+03  14.585  < 2e-16 ***
## stateAL                     6.366e+03  8.204e+03   0.776 0.437804
## stateAR                     7.475e+03  8.301e+03   0.900 0.367896
```

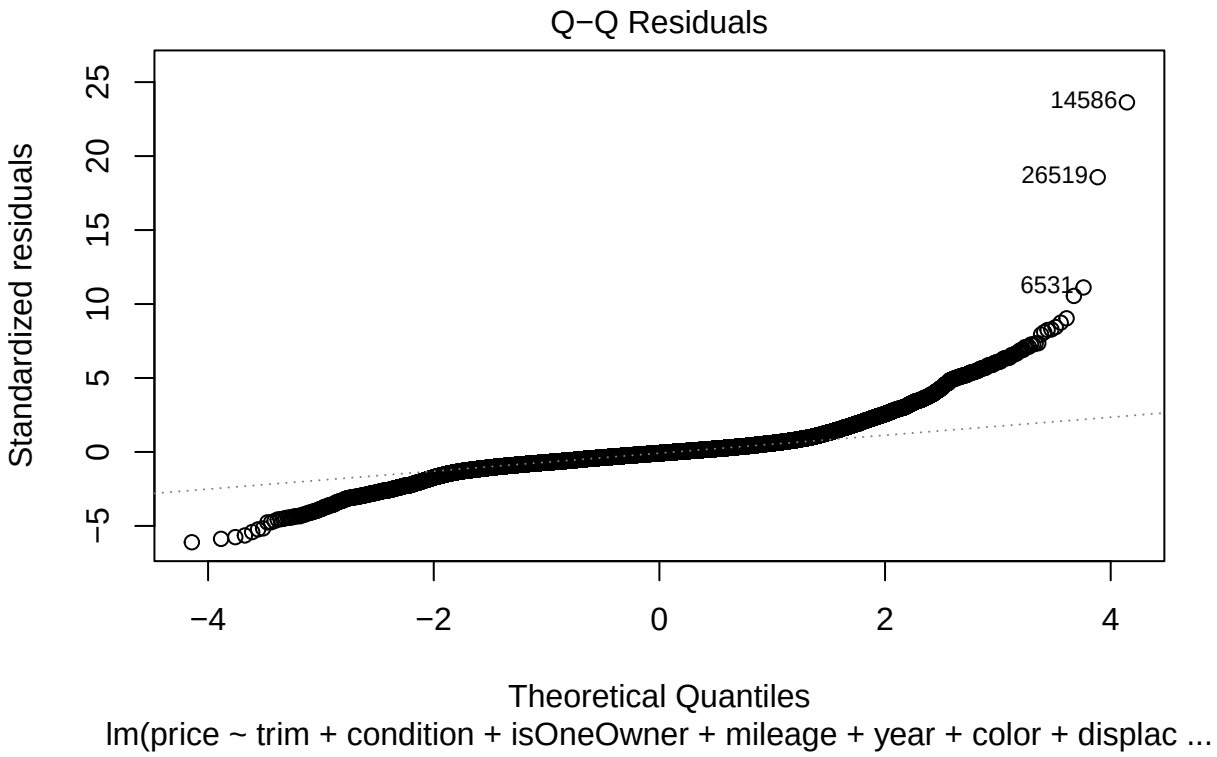
```
## stateAZ                  8.085e+03  8.201e+03   0.986 0.324173
## stateCA                  7.077e+03  8.182e+03   0.865 0.387034
## stateCO                  7.939e+03  8.199e+03   0.968 0.332885
## stateCT                  6.573e+03  8.203e+03   0.801 0.422932
## stateDC                 -7.574e+03  1.056e+04  -0.717 0.473085
## stateDE                  9.059e+03  8.264e+03   1.096 0.273012
## stateFL                  6.662e+03  8.182e+03   0.814 0.415538
## stateGA                  6.558e+03  8.186e+03   0.801 0.423093
## stateHI                  6.723e+03  8.258e+03   0.814 0.415567
## stateIA                  8.349e+03  8.393e+03   0.995 0.319821
## stateID                  1.163e+04  8.528e+03   1.363 0.172743
## stateIL                  7.422e+03  8.186e+03   0.907 0.364603
## stateIN                  6.316e+03  8.222e+03   0.768 0.442421
## stateKS                  7.756e+03  8.277e+03   0.937 0.348761
## stateKY                  1.011e+04  8.220e+03   1.230 0.218887
## stateLA                  8.855e+03  8.235e+03   1.075 0.282216
## stateMA                  7.260e+03  8.190e+03   0.886 0.375398
## stateMD                  7.618e+03  8.192e+03   0.930 0.352442
## stateME                  5.265e+03  8.514e+03   0.618 0.536283
## stateMI                  7.128e+03  8.217e+03   0.868 0.385659
## stateMN                  8.428e+03  8.215e+03   1.026 0.304932
## stateMO                  8.704e+03  8.202e+03   1.061 0.288653
## stateMS                  8.728e+03  8.243e+03   1.059 0.289649
## stateMT                  1.365e+04  8.894e+03   1.535 0.124806
## stateNC                  7.835e+03  8.190e+03   0.957 0.338702
## stateND                  9.940e+03  9.680e+03   1.027 0.304518
## stateNE                  9.993e+03  8.579e+03   1.165 0.244116
## stateNH                  9.080e+03  8.236e+03   1.102 0.270291
## stateNJ                  6.879e+03  8.184e+03   0.841 0.400599
## stateNM                  7.447e+03  8.418e+03   0.885 0.376369
## stateNV                  9.562e+03  8.202e+03   1.166 0.243676
## stateNY                  5.872e+03  8.183e+03   0.718 0.473026
## stateOH                  6.637e+03  8.193e+03   0.810 0.417886
## stateOK                  6.678e+03  8.224e+03   0.812 0.416796
## stateON                  1.255e+04  1.056e+04   1.188 0.234765
## stateOR                  8.463e+03  8.232e+03   1.028 0.303918
## statePA                  7.332e+03  8.191e+03   0.895 0.370725
## stateRI                  6.579e+03  8.333e+03   0.789 0.429840
## stateSC                  8.099e+03  8.213e+03   0.986 0.324074
## stateSD                  2.655e+04  1.157e+04   2.295 0.021733 *
## stateTN                  6.698e+03  8.199e+03   0.817 0.414011
## stateTX                  7.736e+03  8.183e+03   0.945 0.344513
## stateunsp                1.226e+04  1.417e+04   0.866 0.386693
## stateUT                  1.006e+04  8.243e+03   1.221 0.222163
## stateVA                  7.116e+03  8.188e+03   0.869 0.384797
## stateWA                  8.968e+03  8.206e+03   1.093 0.274419
## stateWI                  7.977e+03  8.234e+03   0.969 0.332661
## stateWV                  8.268e+03  8.388e+03   0.986 0.324306
## stateWY                  1.984e+03  1.156e+04   0.172 0.863761
## regionESC                       NA         NA      NA       NA
## regionMid                       NA         NA      NA       NA
## regionMtn                       NA         NA      NA       NA
## regionNew                       NA         NA      NA       NA
## regionPac                       NA         NA      NA       NA
```

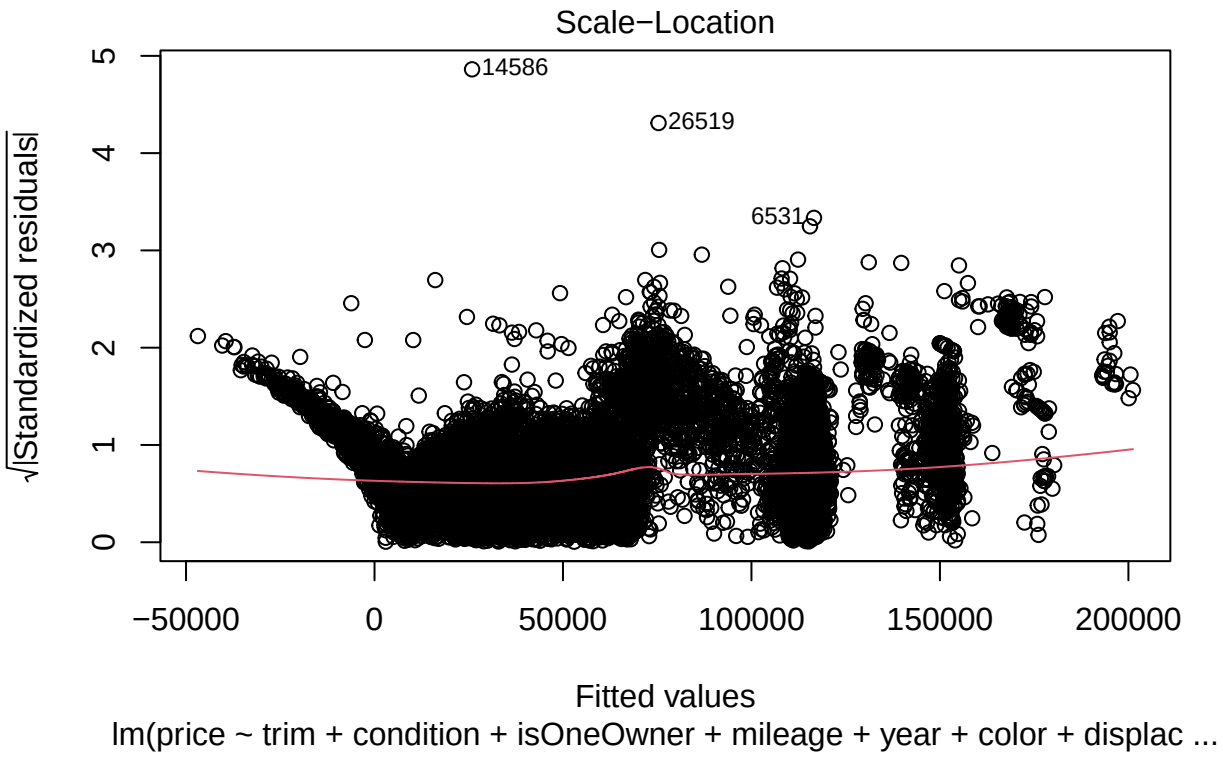
```
## regionSoA                          NA         NA      NA       NA
## regionunsp                         NA         NA      NA       NA
## regionWNC                          NA         NA      NA       NA
## regionWSC                          NA         NA      NA       NA
## soundSystemBang Olufsen    -1.522e+03  8.233e+03  -0.185 0.853352
## soundSystemBose            -9.304e+03  8.193e+03  -1.136 0.256143
## soundSystemBoston Acoustic -1.202e+04  1.417e+04  -0.848 0.396180
## soundSystemHarman Kardon   -9.387e+03  8.187e+03  -1.147 0.251569
## soundSystemPremium         -7.230e+03  8.185e+03  -0.883 0.377059
## soundSystemunsp            -6.913e+03  8.185e+03  -0.845 0.398363
## wheelTypeChrome            -2.751e+02  1.304e+03  -0.211 0.832883
## wheelTypePremium           -4.943e+02  5.841e+02  -0.846 0.397436
## wheelTypeSteel              1.643e+04  1.663e+03   9.877  < 2e-16 ***
## wheelTypeunsp               5.467e+01  1.689e+02   0.324 0.746153
## featureCount               -7.875e+00  3.026e+00  -2.603 0.009258 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11560 on 29194 degrees of freedom
##   (174 observations deleted due to missingness)
## Multiple R-squared:  0.9327, Adjusted R-squared:  0.9325
## F-statistic:  4171 on 97 and 29194 DF,  p-value: < 2.2e-16
```

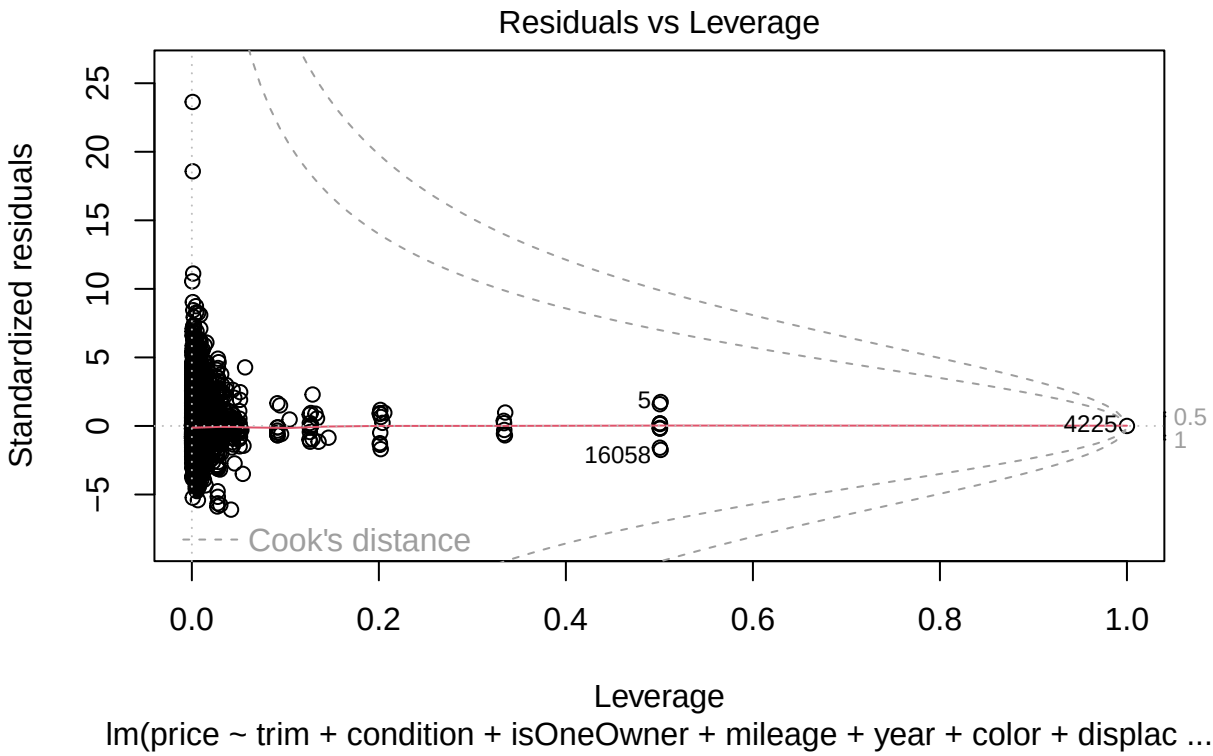
Based on our results we can see that there are quite a few significant variables, however most of these are categorical and are related to one specific column, that being the trim (model) of the vehicle. The most significant predictors appear to be the trim, mileage, new/used status, year, and displacement. Also notably, featureCount is also quite significant but not to the same degree as the ones listed before this. This makes a lot of sense as in the real world these are often what people use as the primary indicators on condition and the usable life left of a car, then additional featureCount being an added value secondary to the initial evaluation of the more critical features.

Some of these variables are probably highly correlated. Especially mileage and year as typically the older a car is the more miles it'll have.




Q−Q Residuals
Standardized residuals
25
20
15
10
5
0
−5
14586
26519
6531
−4
−2
0
2
4
Theoretical Quantiles
lm(price ~ trim + condition + isOneOwner + mileage + year + color + displac ...

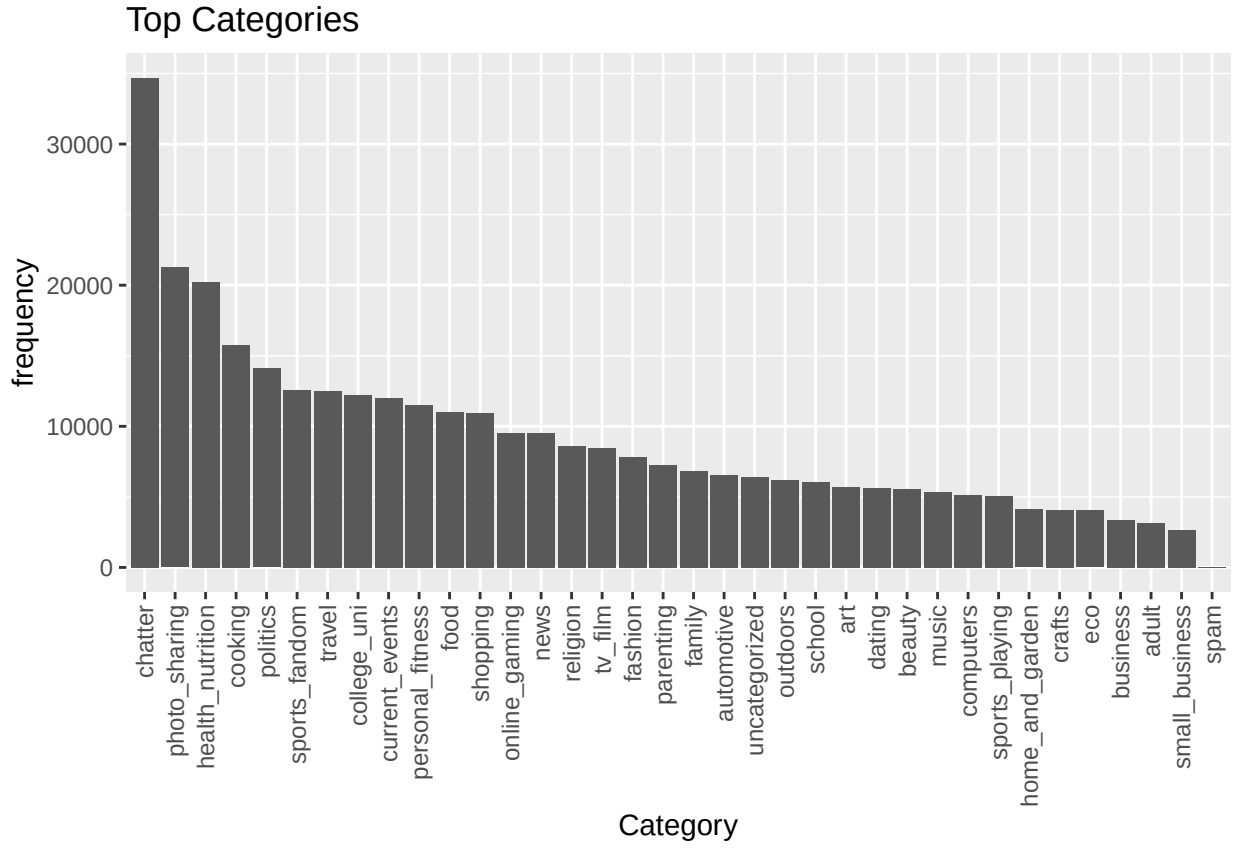

Scale−Location
14586
26519
6531
√|Standardized residuals|
0
1
2
3
4
5
−50000
0
50000
100000
150000
200000
Fitted values
lm(price ~ trim + condition + isOneOwner + mileage + year + color + displac ...



Especially based on the Q-Q residual plot we can see that currently we are not describing the upper most and lower most predictions very well, with our more central predictions seeming to be the most accurate which makes sense as value of the lowest and highest values probably being influenced by something that isn't being captured effectively by our data.

Data Set #2 - Twitter Posting

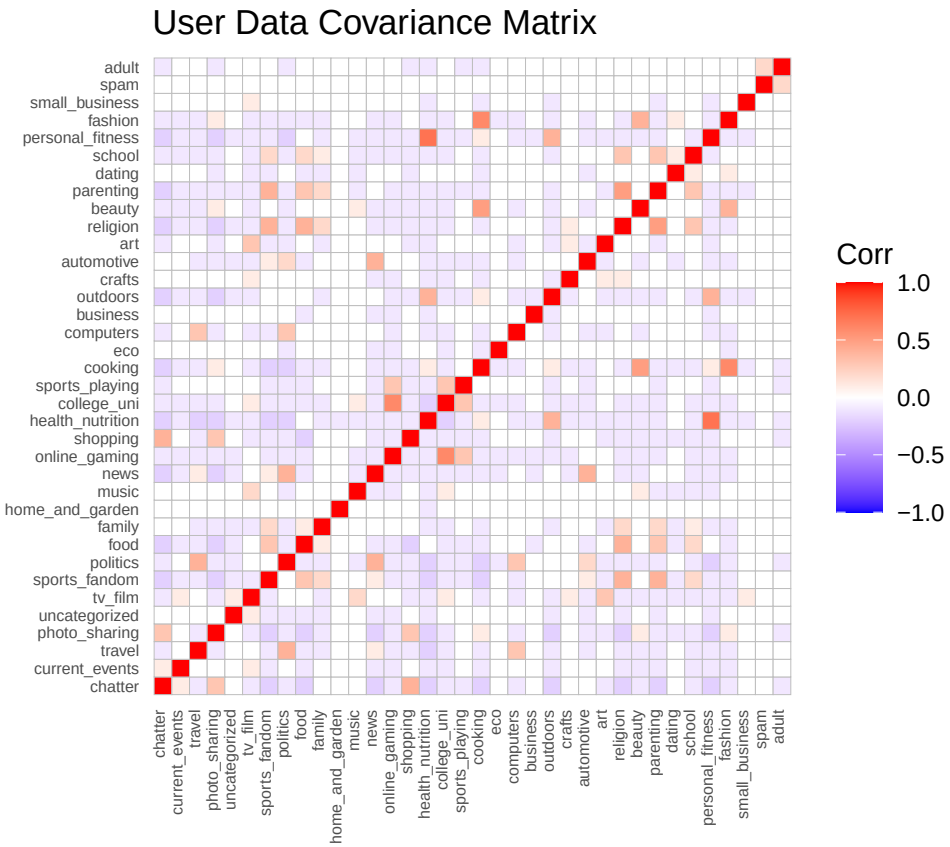
We can see that the anti-botting has done pretty well as explained in the description with a very small amount of posts being spam/adult content. To start I'm going to go ahead and scale every row so that the sum of every row equals 1. This way when going to our next step we aren't disproportionately weighting any features when training our classification model.



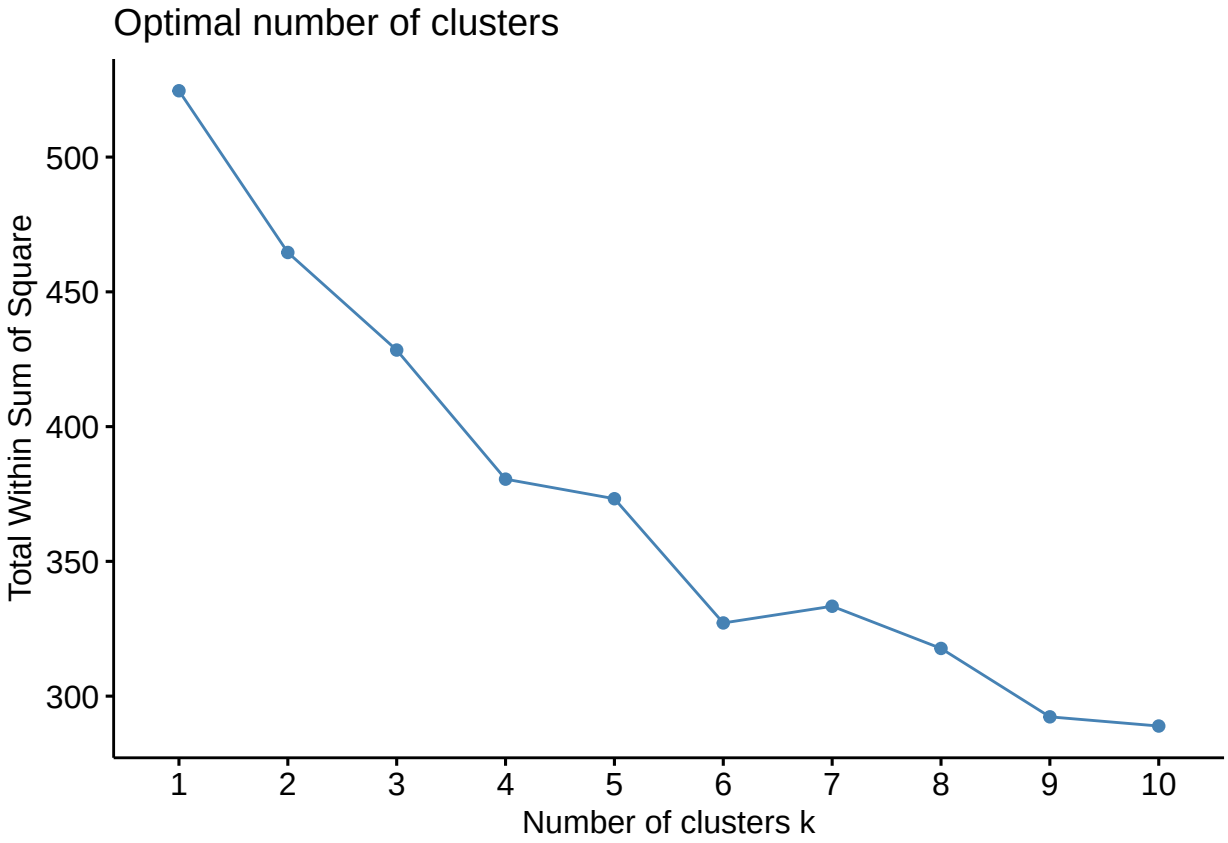

```
## # A tibble: 7,882 x 37
##    user_ID   chatter current_events travel photo_sharing uncategorized tv_film
##    <fct>       <dbl>          <dbl>  <dbl>         <dbl>         <dbl>   <dbl>
##  1 hmjoe4g3k  0.0328         0      0.0328        0.0328        0.0328  0.0164
##  2 clk1m5w8s  0.1            0.1    0.0667        0.0333        0.0333  0.0333
##  3 jcsovtak3  0.128          0.0638 0.0851        0.0638        0.0213  0.106
##  4 3oeb4hiln  0.0476         0.238  0.0952        0.0952        0       0.0476
##  5 fd75x1vgk  0.167          0.0667 0             0.2           0.0333  0
##  6 h6nvj91yp  0.176          0.118  0.0588        0.206         0       0.0294
##  7 ma7kfewxq  0.0263         0.0526 0.184         0.0263        0       0.0263
##  8 u48d61ztj  0.1            0.06   0.06          0.12          0.02    0.02
##  9 y2g68vhkf  0.08           0.0267 0             0.0133        0       0
## 10 n467yj1st  0.0575         0.0230 0.0460        0.0460        0       0.0575
## # i 7,872 more rows
## # i 30 more variables: sports_fandom <dbl>, politics <dbl>, food <dbl>,
## #   family <dbl>, home_and_garden <dbl>, music <dbl>, news <dbl>,
## #   online_gaming <dbl>, shopping <dbl>, health_nutrition <dbl>,
## #   college_uni <dbl>, sports_playing <dbl>, cooking <dbl>, eco <dbl>,
## #   computers <dbl>, business <dbl>, outdoors <dbl>, crafts <dbl>,
```

```
## #   automotive <dbl>, art <dbl>, religion <dbl>, beauty <dbl>, ...
```

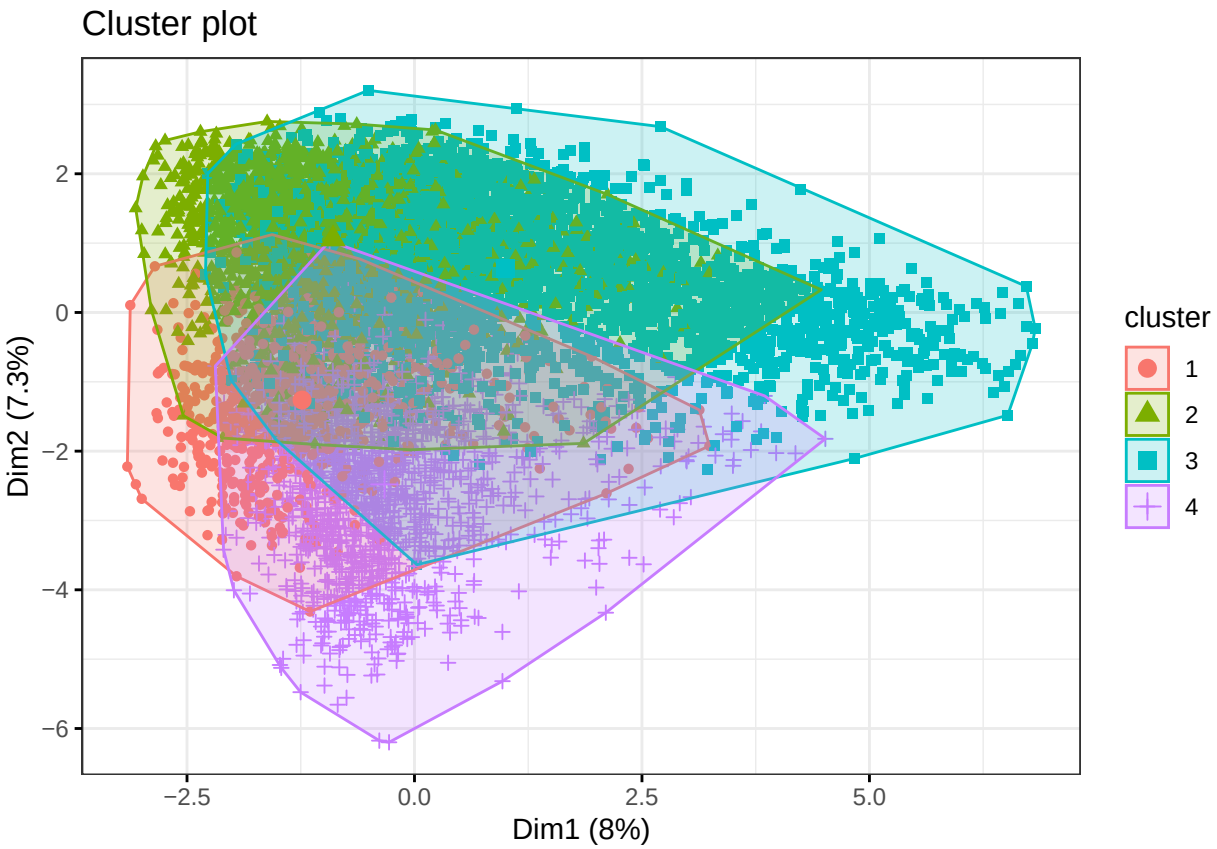
As part of our continued exploratory analysis, we can see that a lot of our categories have covariance with one or more other categories making me think that we'll be fairly capable of summarizing our users into distinct groupings to give us a better idea of some of our user demographics.

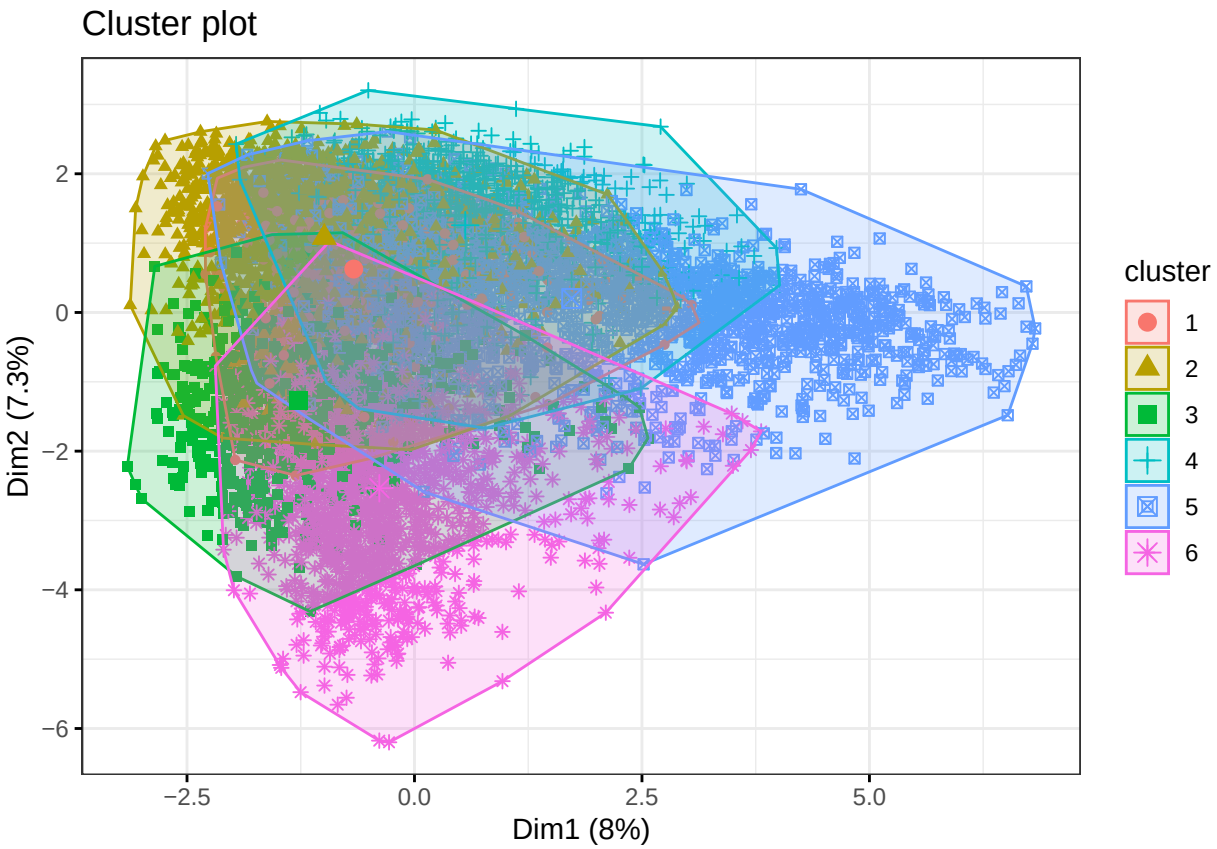


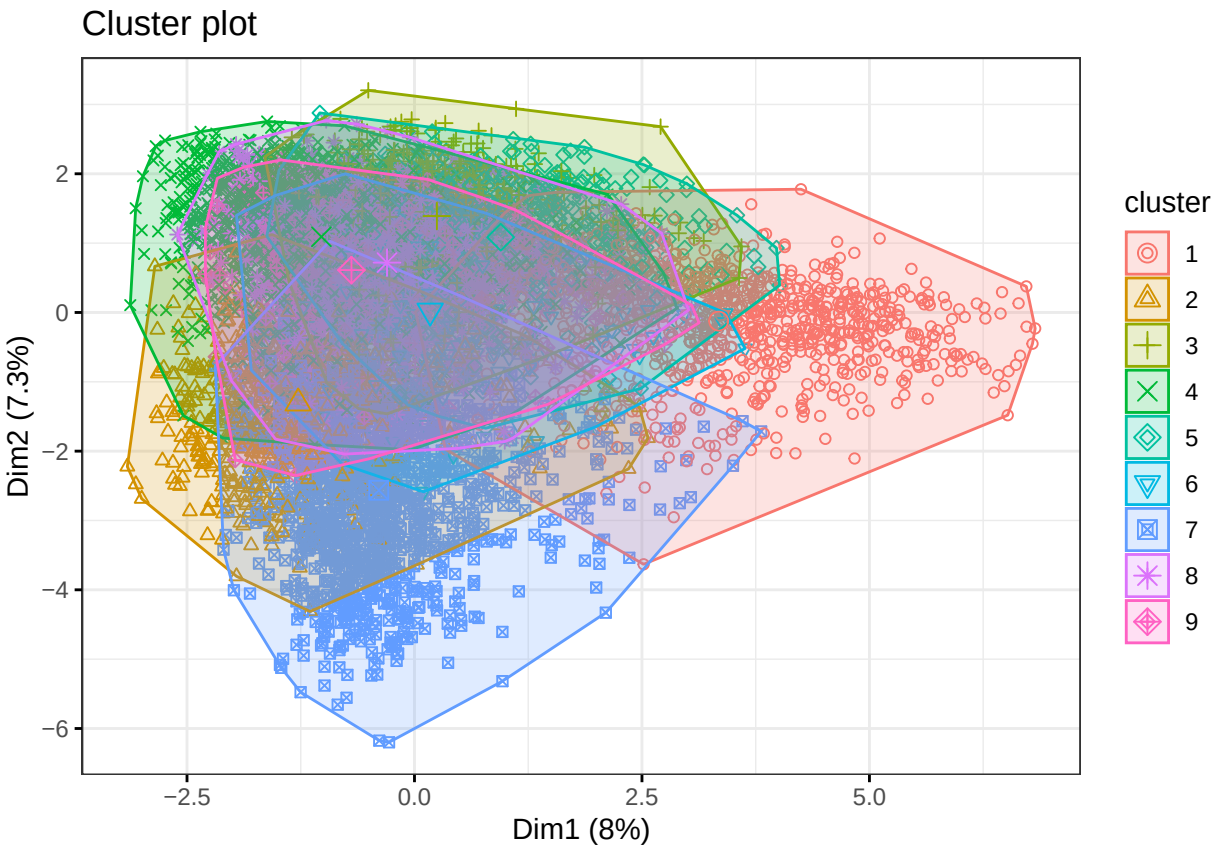
Looking at our number of clusters analysis we can take a shot and say that the optimal number is probably 4, 6, or 9. I think it'll be worth it to go ahead and do multiple models corresponding with each of these predictably optimal number of clusters.



Now we can go ahead and train our various models targeting 4, 6, and 9 centers.


Cluster plot
Dim2 (7.3%)
2
0
−2
−4
−6
−2.5
0.0
2.5
5.0
Dim1 (8%)
cluster
1
2
3
4

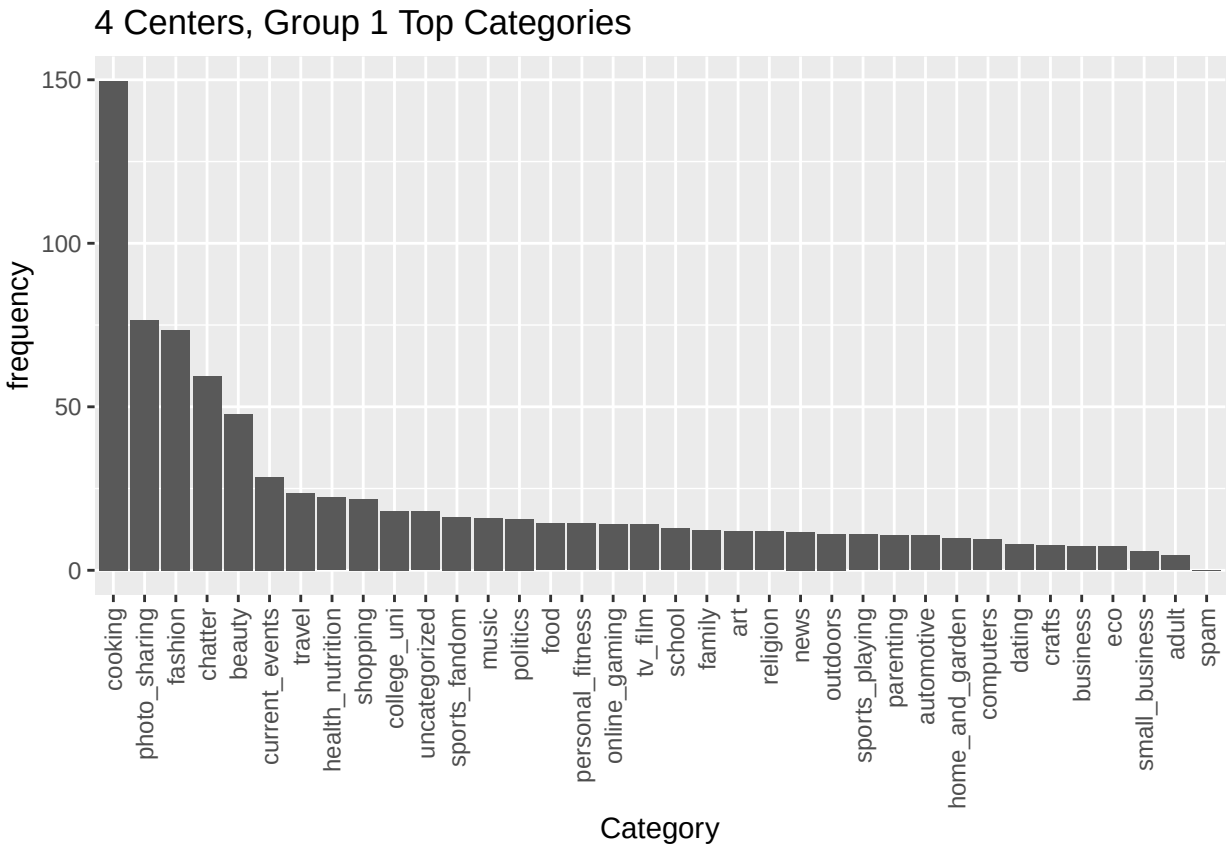

Cluster plot
Dim2 (7.3%)
2
0
−2
−4
−6
−2.5
0.0
2.5
5.0
Dim1 (8%)
cluster
1
2
3
4
5
6

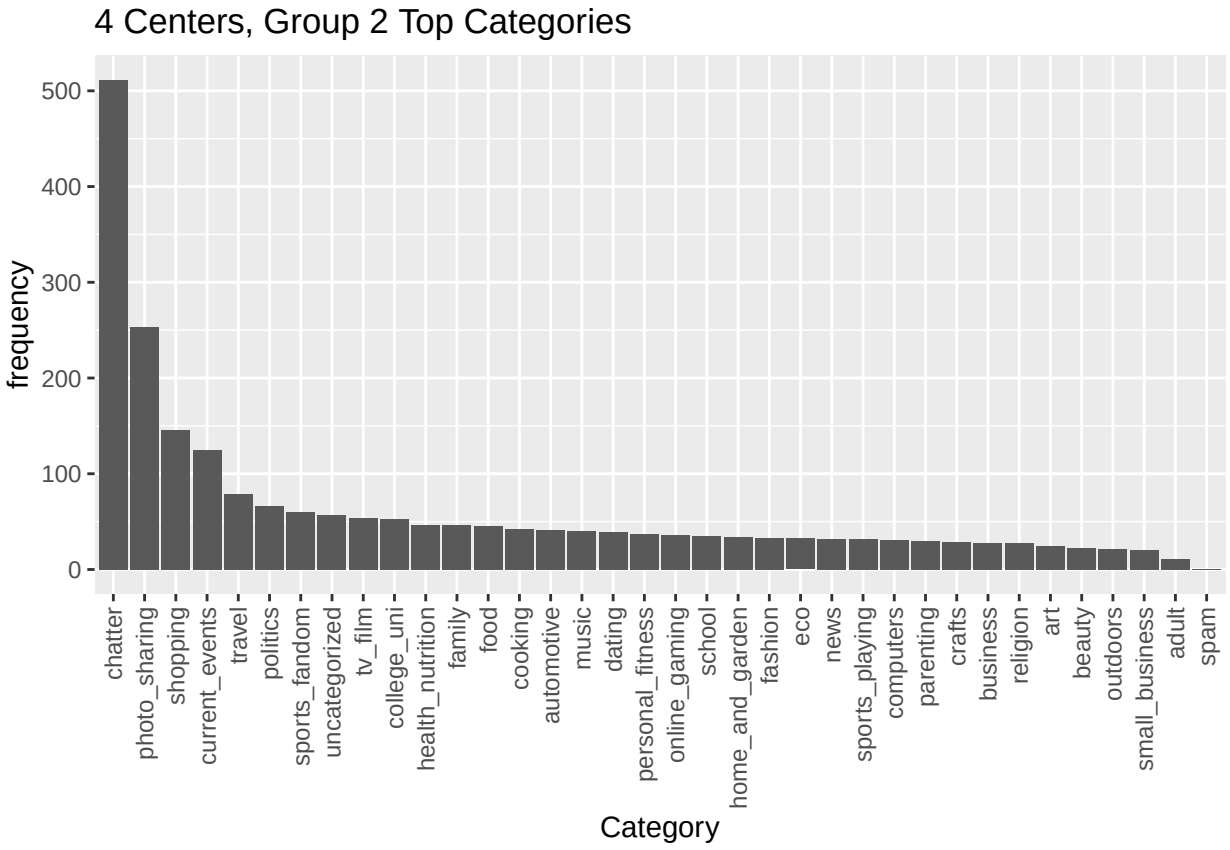


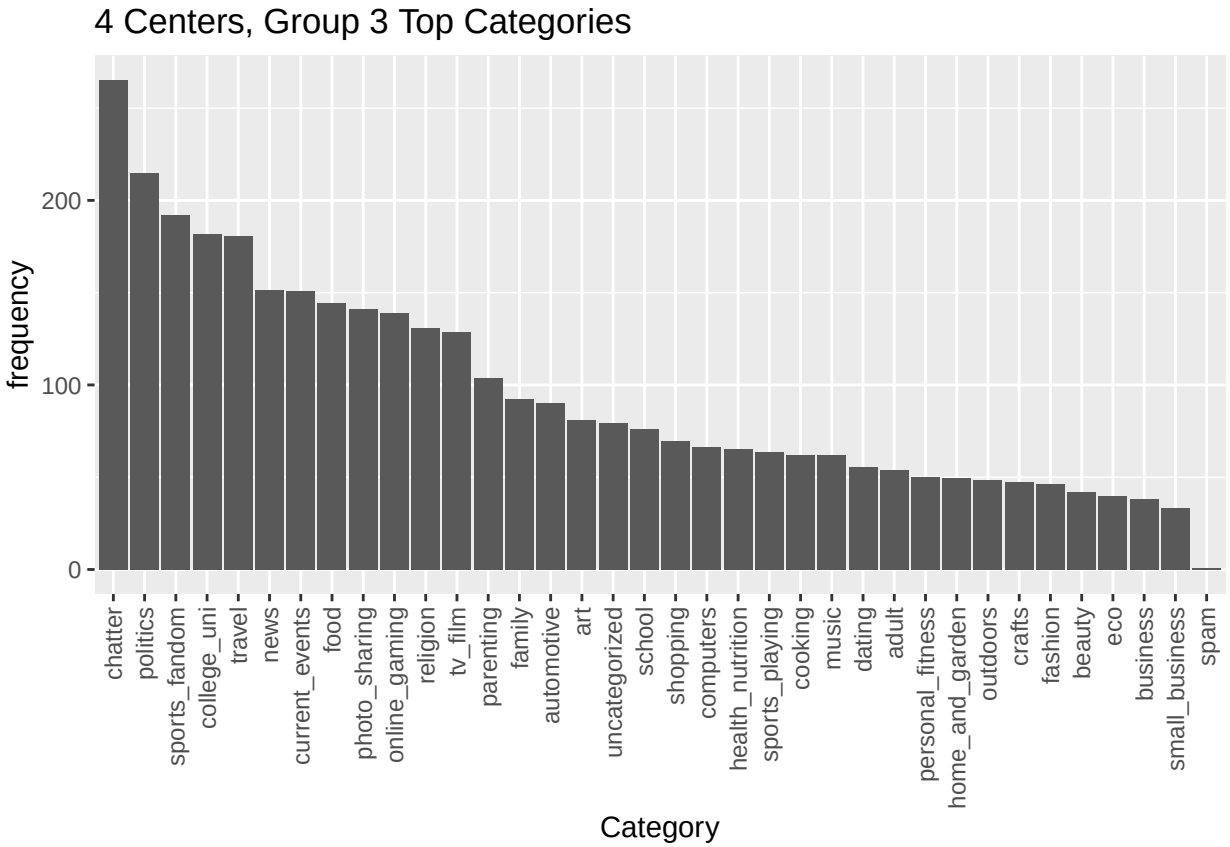
Initially looking at these plots they may look like they are a mess, but since we can only look at 2 of the dimensions at a time, we are missing a lot of information that we can't really visually see so lets see if we can better summarize these groups by a summary of their groups.

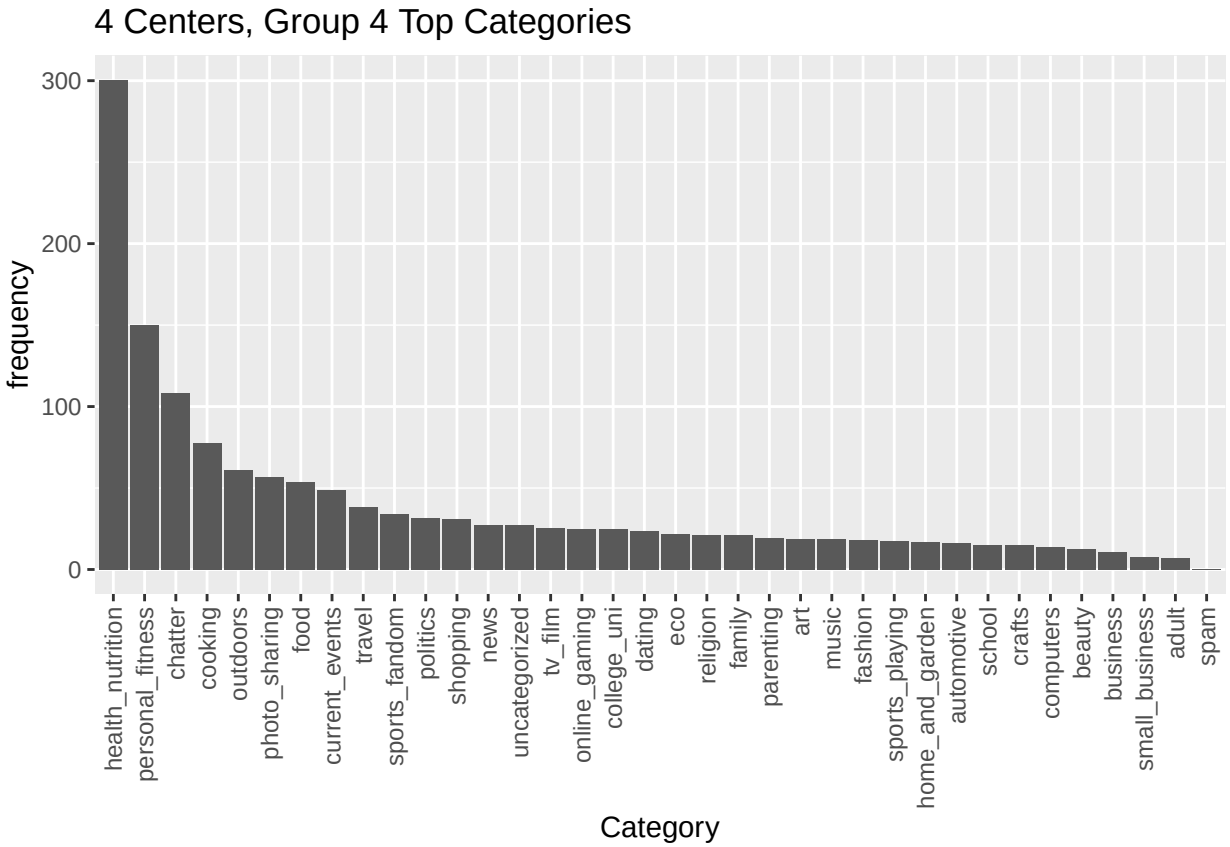
To start we're going to go ahead and summarize our 4 center model groups. Our groups appear to be somewhat distinct.

- Group 1 - "Women" This group definition is based on the fact that they are primarily interested in cooking, fashion, and beauty.
- Group 2 - "Chatter/Photo Sharing" A catch-all group where the user doesn't appear to really fall into any specific group other than chatter and photo sharing.
- Group 3 - "College Students" This is based on the fact that they are very involved in general chatter, but are also equally interested in politics, college/uni, travel, news, current, events, and food. All pretty indicative activities of college students.
- Group 4 - "Health Nuts" These people are very interested in health/nutrition, personal fitness, cooking, and the outdoors.



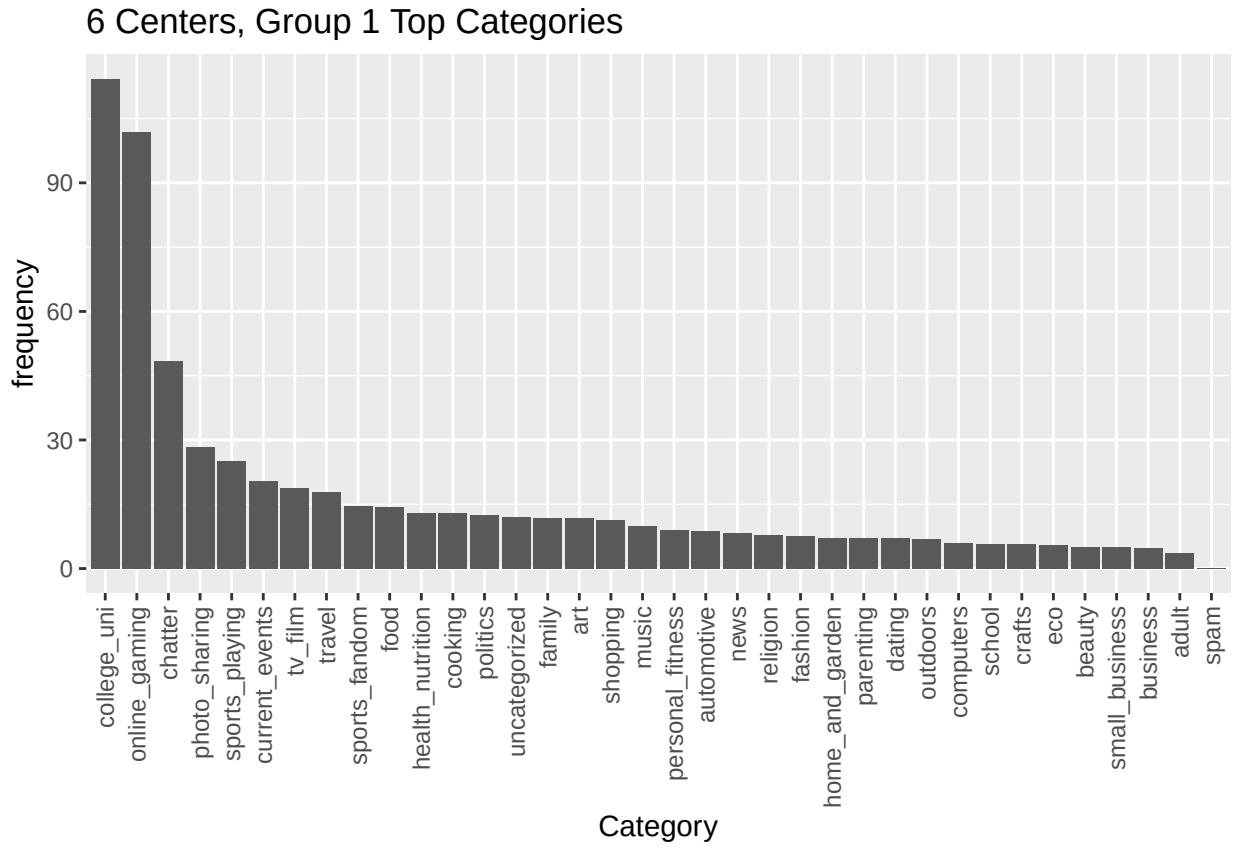

4 Centers, Group 2 Top Categories
frequency
500
400
300
200
100
0
chatter
photo_sharing
shopping
current_events
travel
politics
sports_fandom
uncategorized
tv_film
college_uni
health_nutrition
family
food
cooking
automotive
music
dating
personal_fitness
online_gaming
school
home_and_garden
fashion
eco
news
sports_playing
computers
parenting
crafts
business
religion
art
beauty
outdoors
small_business
adult
spam
Category

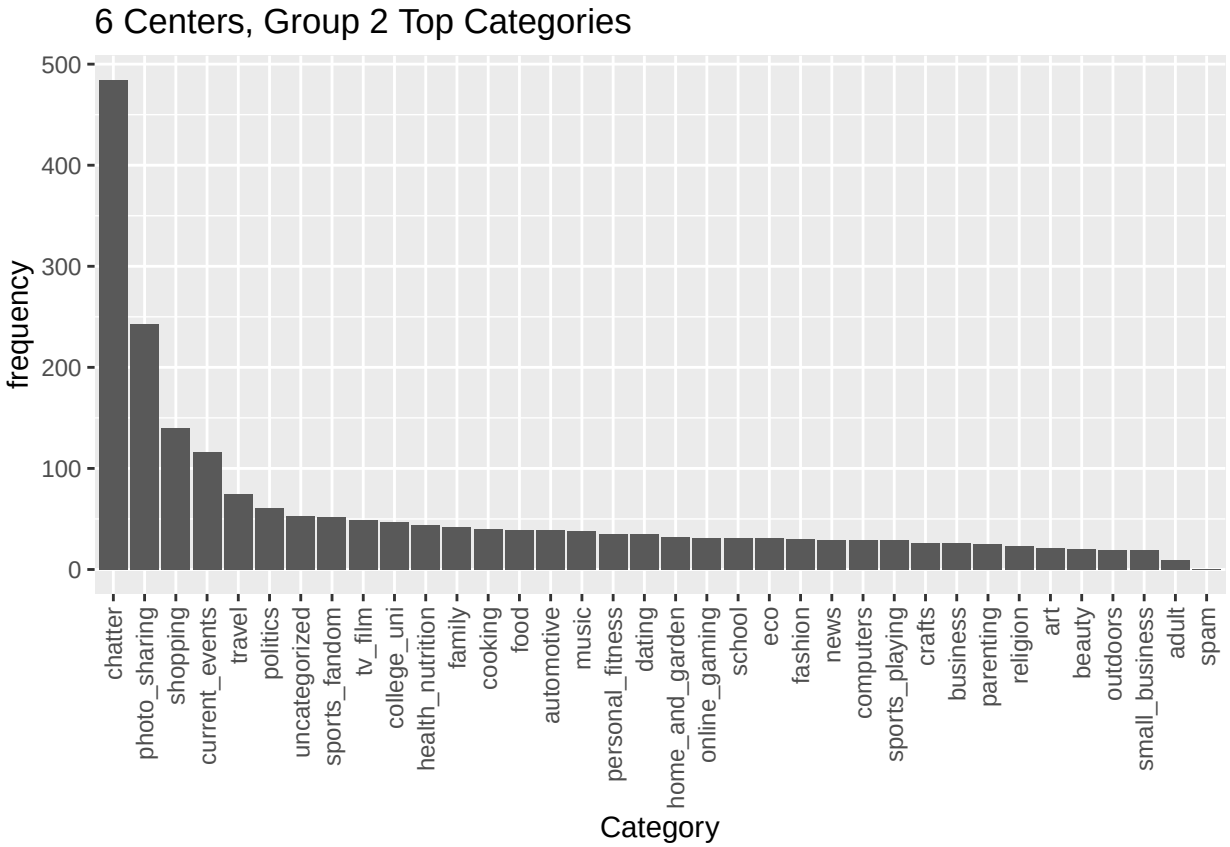

4 Centers, Group 3 Top Categories
frequency
200
100
0
chatter
politics
sports_fandom
college_uni
travel
news
current_events
food
photo_sharing
online_gaming
religion
tv_film
parenting
family
automotive
art
uncategorized
school
shopping
computers
health_nutrition
sports_playing
cooking
music
dating
adult
personal_fitness
home_and_garden
outdoors
crafts
fashion
beauty
eco
business
small_business
spam
Category

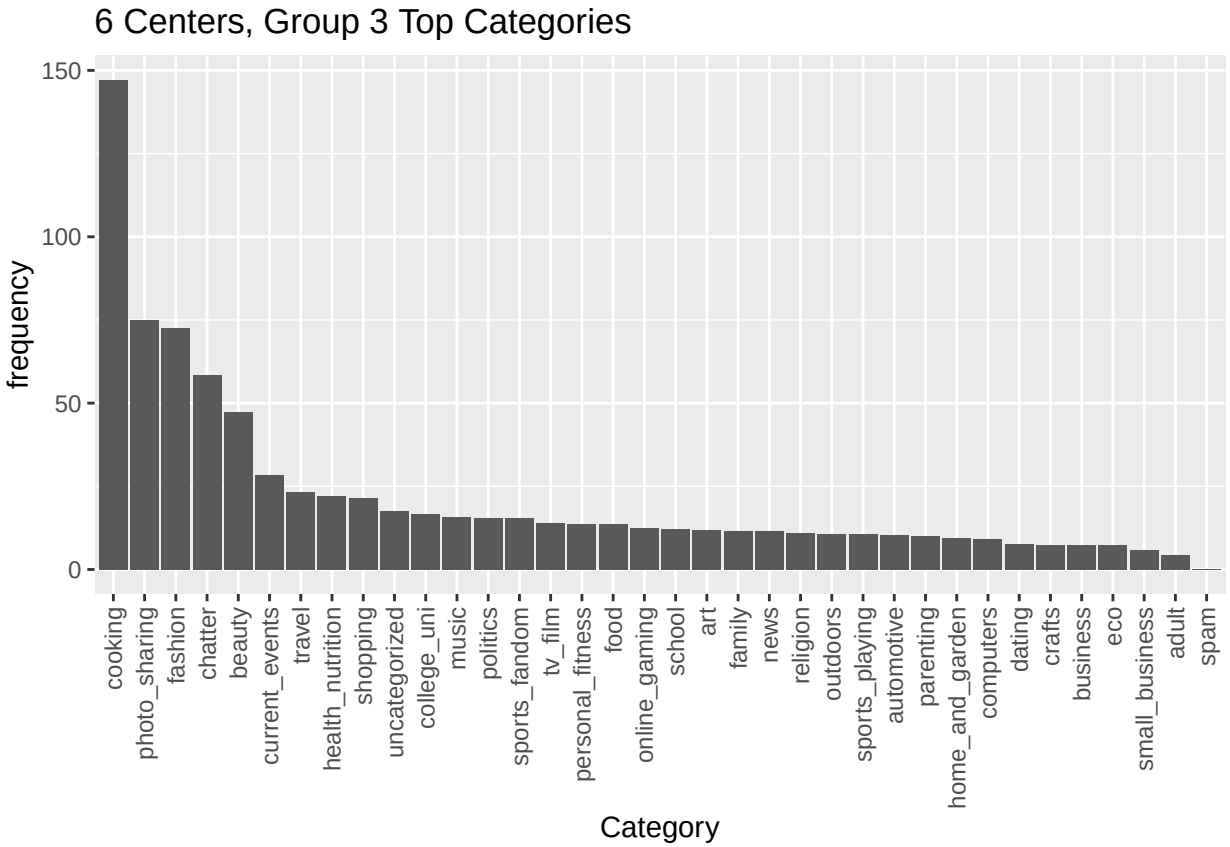

4 Centers, Group 4 Top Categories
frequency
300
200
100
0
health_nutrition
personal_fitness
chatter
cooking
outdoors
photo_sharing
food
current_events
travel
sports_fandom
politics
shopping
news
uncategorized
tv_film
online_gaming
college_uni
dating
eco
religion
family
parenting
art
music
fashion
sports_playing
home_and_garden
automotive
school
crafts
computers
beauty
business
small_business
adult
spam
Category

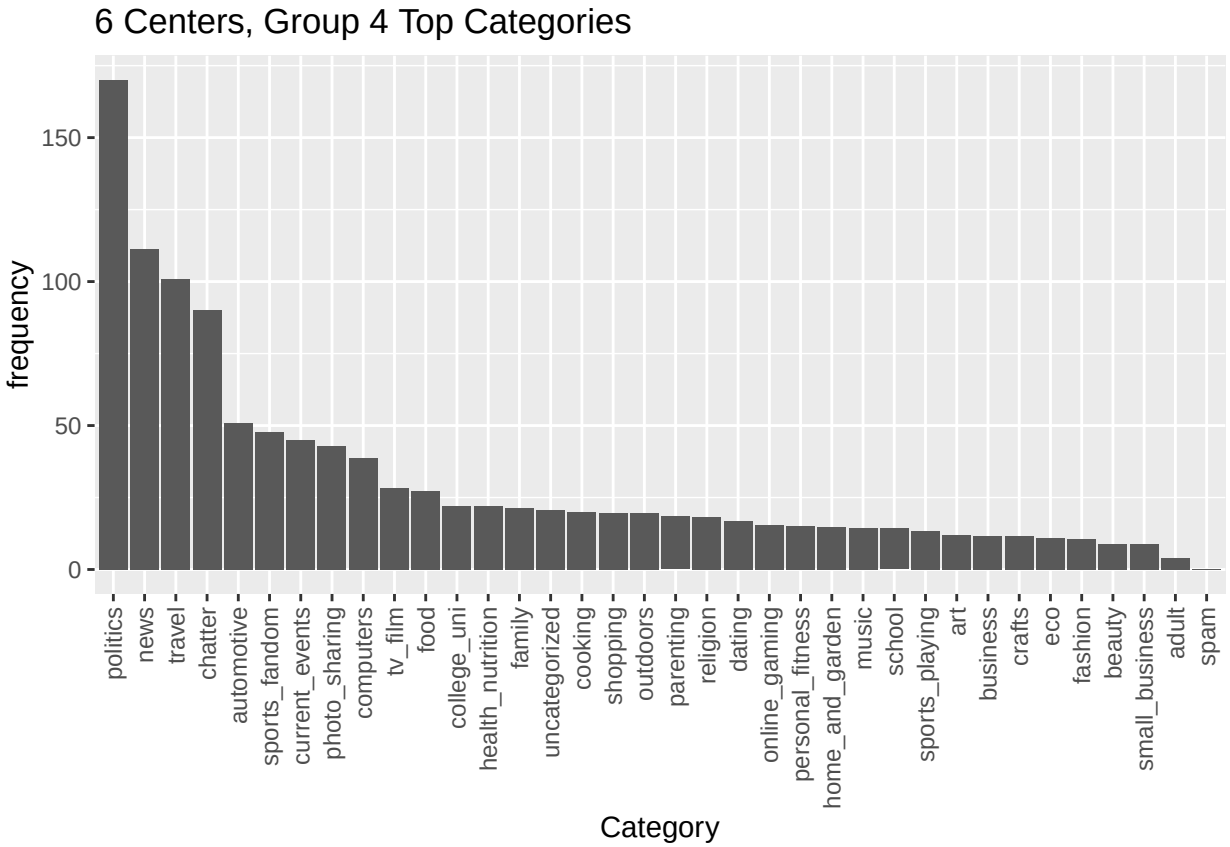
Moving on we're going to go ahead and summarize our 6 center model groups. Once again our groups appear to be somewhat distinct.

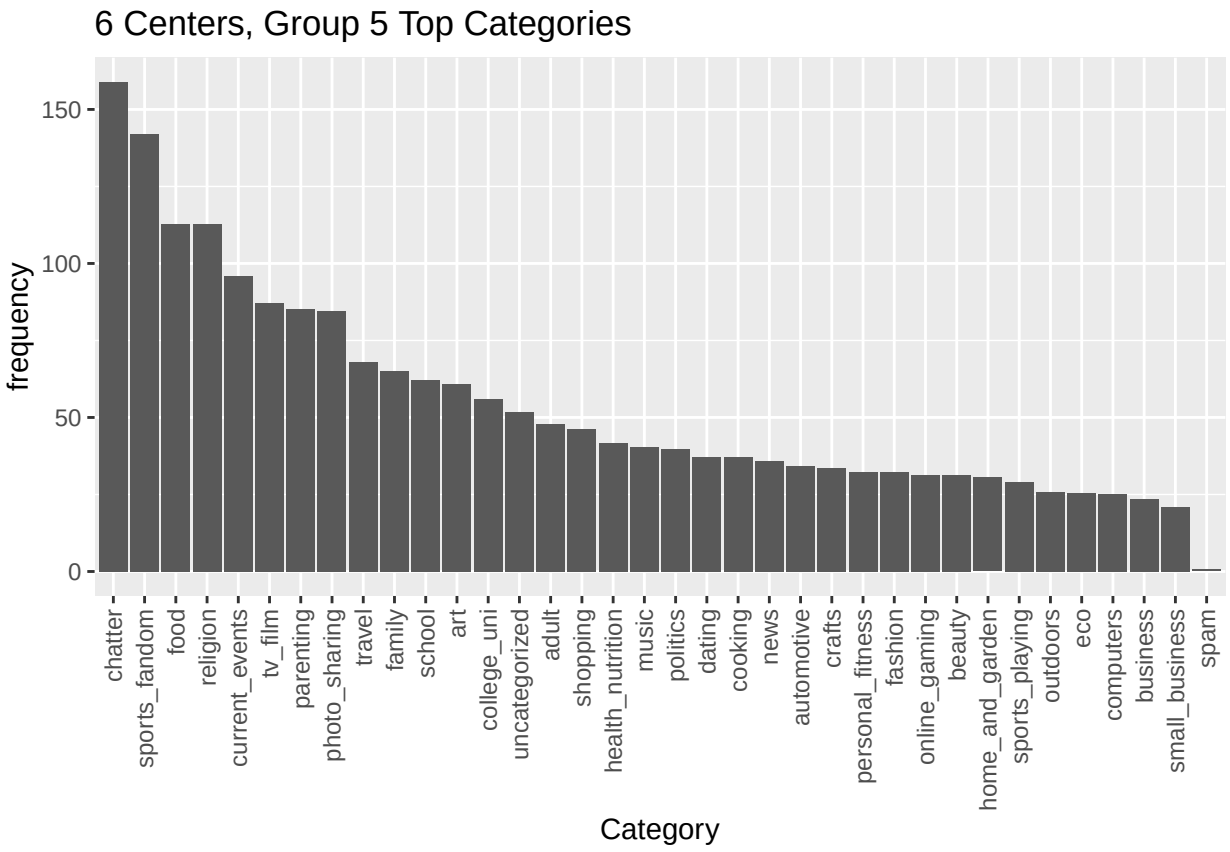
- Group 1 - "College Age Gamers" This group appears to be primarily interested in college/university and online gaming.
- Group 2 - "Chatter/Photo Sharing" A catch-all group where the user doesn't appear to really fall into any specific group other than chatter and photo sharing.
- Group 3 - "Women" This group definition is based on the fact that they are primarily interested in cooking, fashion, and beauty.
- Group 4 - "Worldly Current Event" This group is primarily interested in politics, news, and travel.
- Group 5 - "Media Consumers" This group engages in a lot of chatter but also engages in sports, food, current events, and tv/film.
- Group 6 - "Health Nuts" These people are very interested in health/nutrition, personal fitness, cooking, and the outdoors.

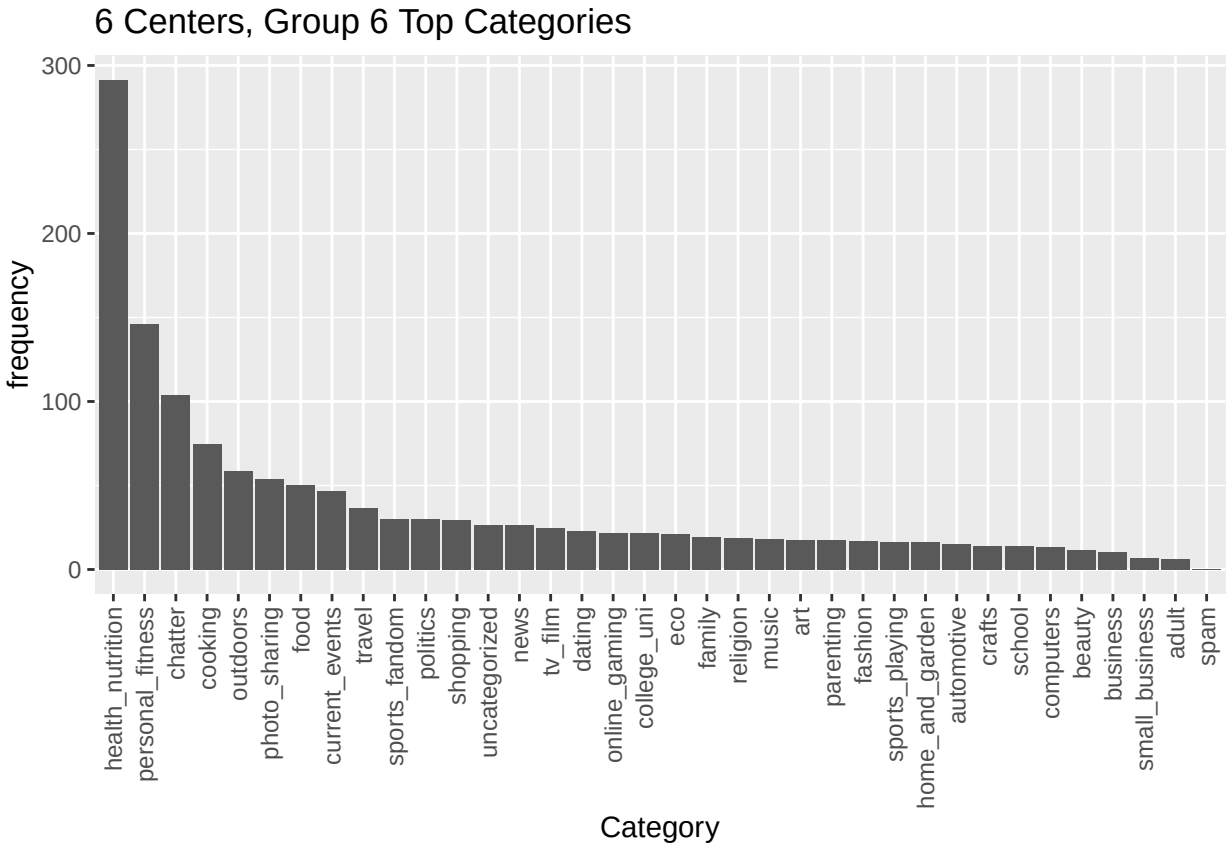



6 Centers, Group 2 Top Categories
frequency
500
400
300
200
100
0
chatter
photo_sharing
shopping
current_events
travel
politics
uncategorized
sports_fandom
tv_film
college_uni
health_nutrition
family
cooking
food
automotive
music
personal_fitness
dating
home_and_garden
online_gaming
school
eco
fashion
news
computers
sports_playing
crafts
business
parenting
religion
art
beauty
outdoors
small_business
adult
spam
Category


6 Centers, Group 3 Top Categories
frequency
150
100
50
0
cooking
photo_sharing
fashion
chatter
beauty
current_events
travel
health_nutrition
shopping
uncategorized
college_uni
music
politics
sports_fandom
tv_film
personal_fitness
food
online_gaming
school
art
family
news
religion
outdoors
sports_playing
automotive
parenting
home_and_garden
computers
dating
crafts
business
eco
small_business
adult
spam
Category

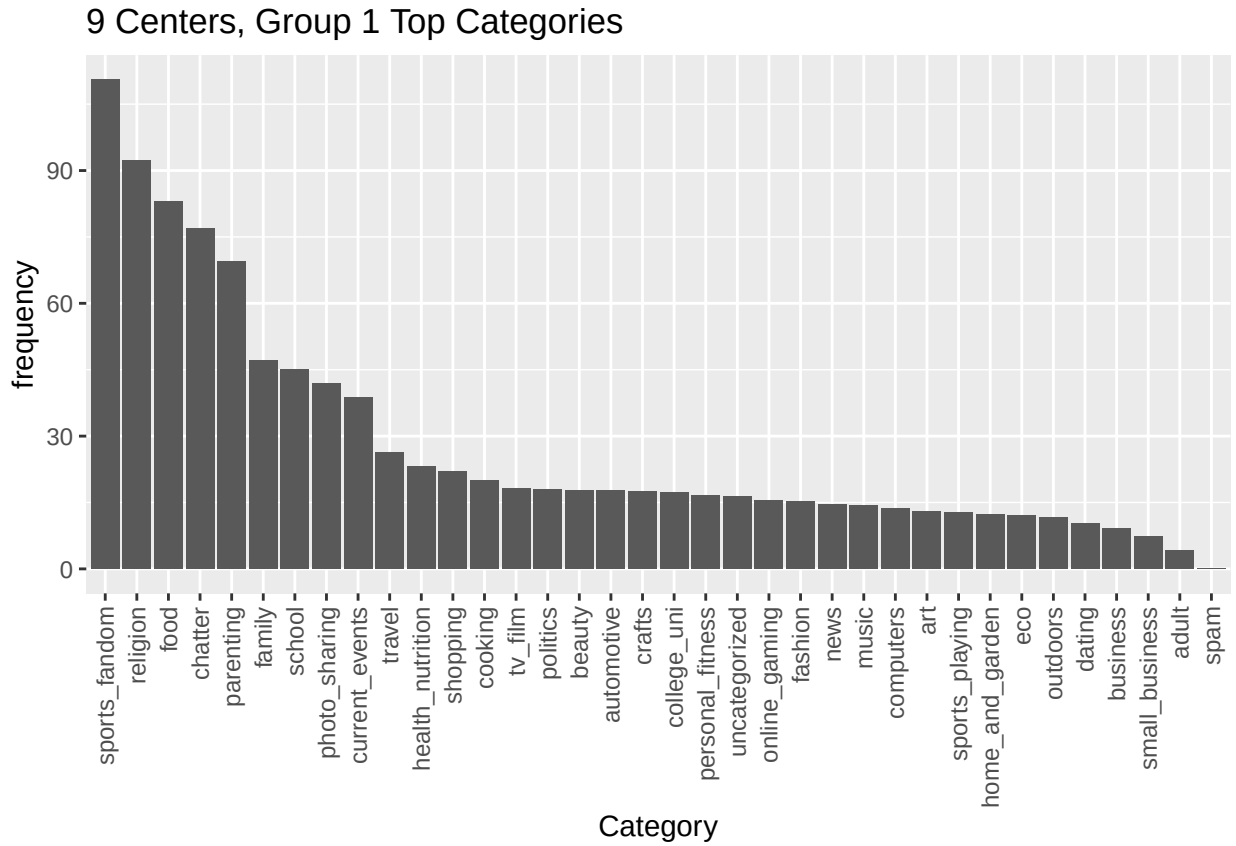

6 Centers, Group 4 Top Categories
frequency
150
100
50
0
politics
news
travel
chatter
automotive
sports_fandom
current_events
photo_sharing
computers
tv_film
food
college_uni
health_nutrition
family
uncategorized
cooking
shopping
outdoors
parenting
religion
dating
online_gaming
personal_fitness
home_and_garden
music
school
sports_playing
art
business
crafts
eco
fashion
beauty
small_business
adult
spam
Category

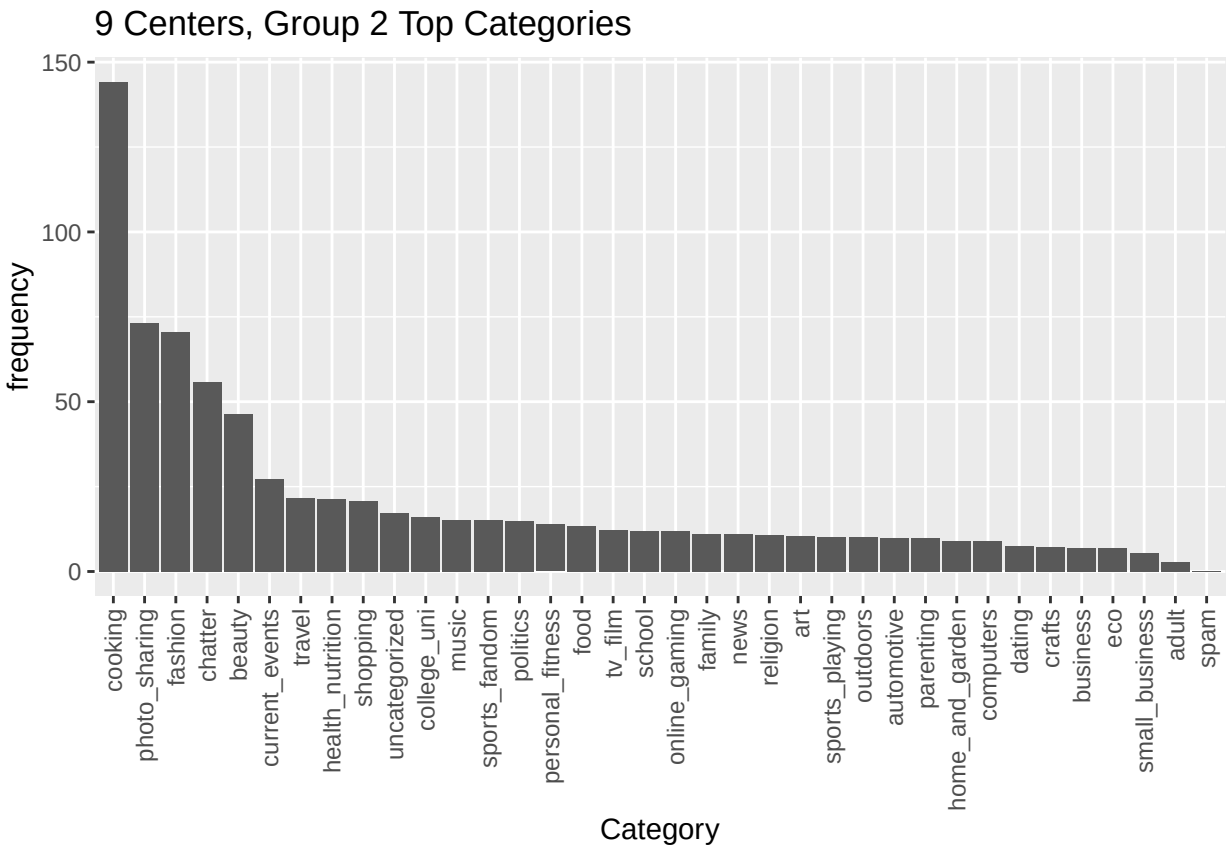

6 Centers, Group 5 Top Categories
frequency
150
100
50
0
chatter
sports_fandom
food
religion
current_events
tv_film
parenting
photo_sharing
travel
family
school
art
college_uni
uncategorized
adult
shopping
health_nutrition
music
politics
dating
cooking
news
automotive
crafts
personal_fitness
fashion
online_gaming
beauty
home_and_garden
sports_playing
outdoors
eco
computers
business
small_business
spam
Category

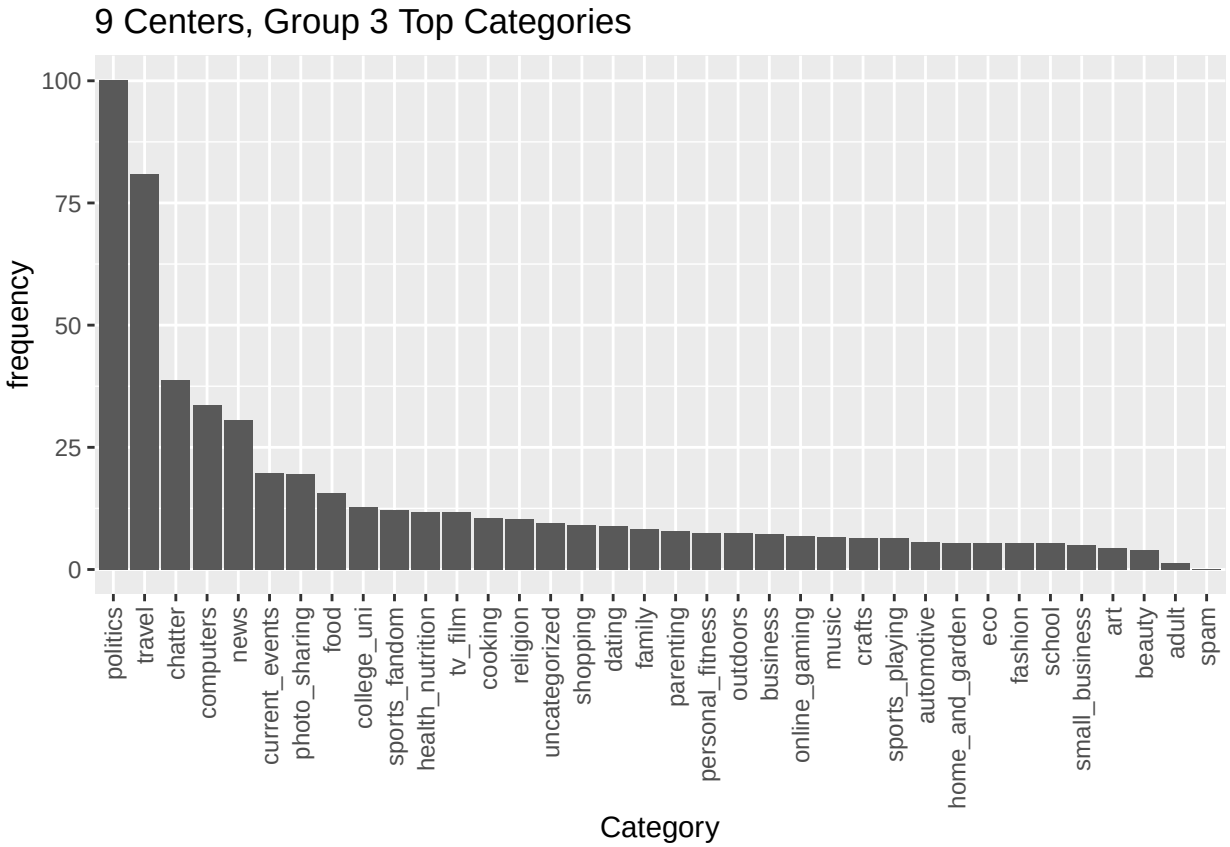

6 Centers, Group 6 Top Categories
frequency
300
200
100
0
health_nutrition
personal_fitness
chatter
cooking
outdoors
photo_sharing
food
current_events
travel
sports_fandom
politics
shopping
uncategorized
news
tv_film
dating
online_gaming
college_uni
eco
family
religion
music
art
parenting
fashion
sports_playing
home_and_garden
automotive
crafts
school
computers
beauty
business
small_business
adult
spam
Category

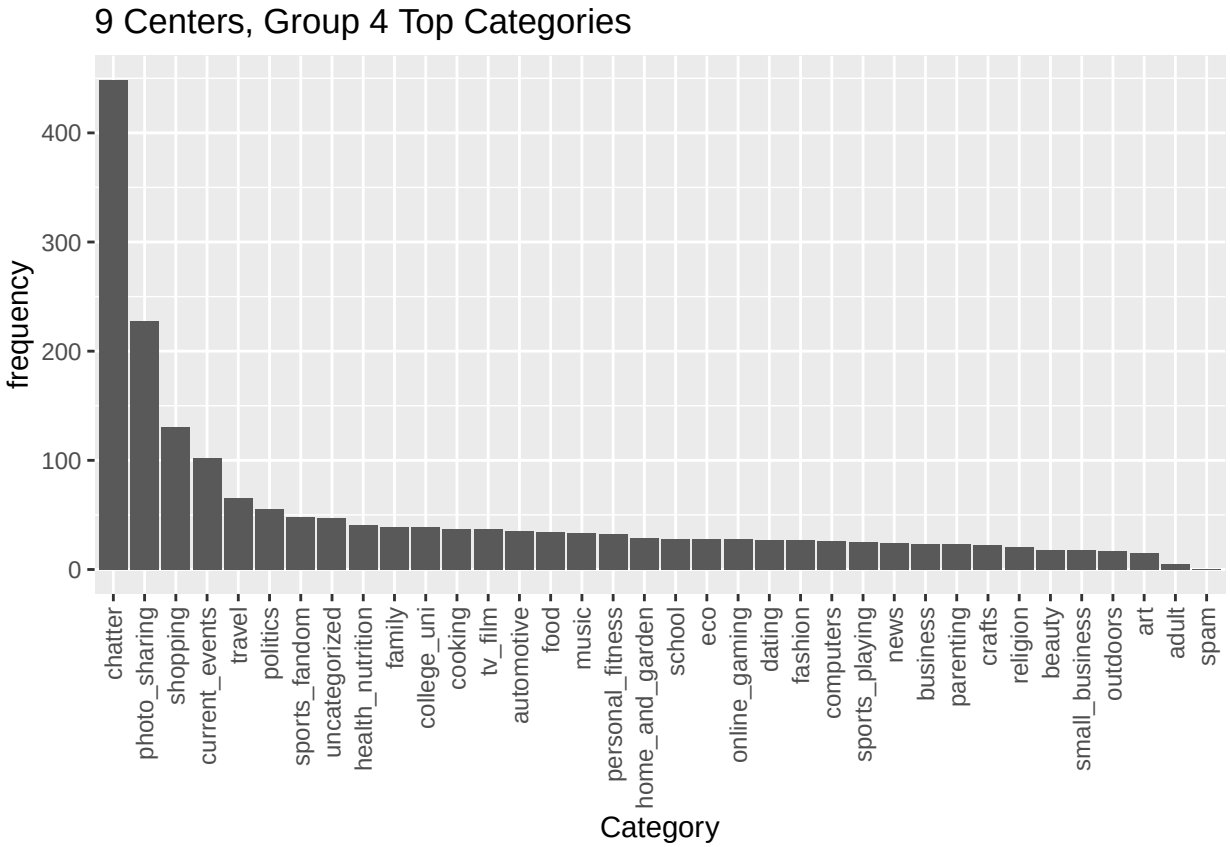
Finally we're going to go ahead and summarize our 9 center model groups. Once again our groups appear to be somewhat distinct.

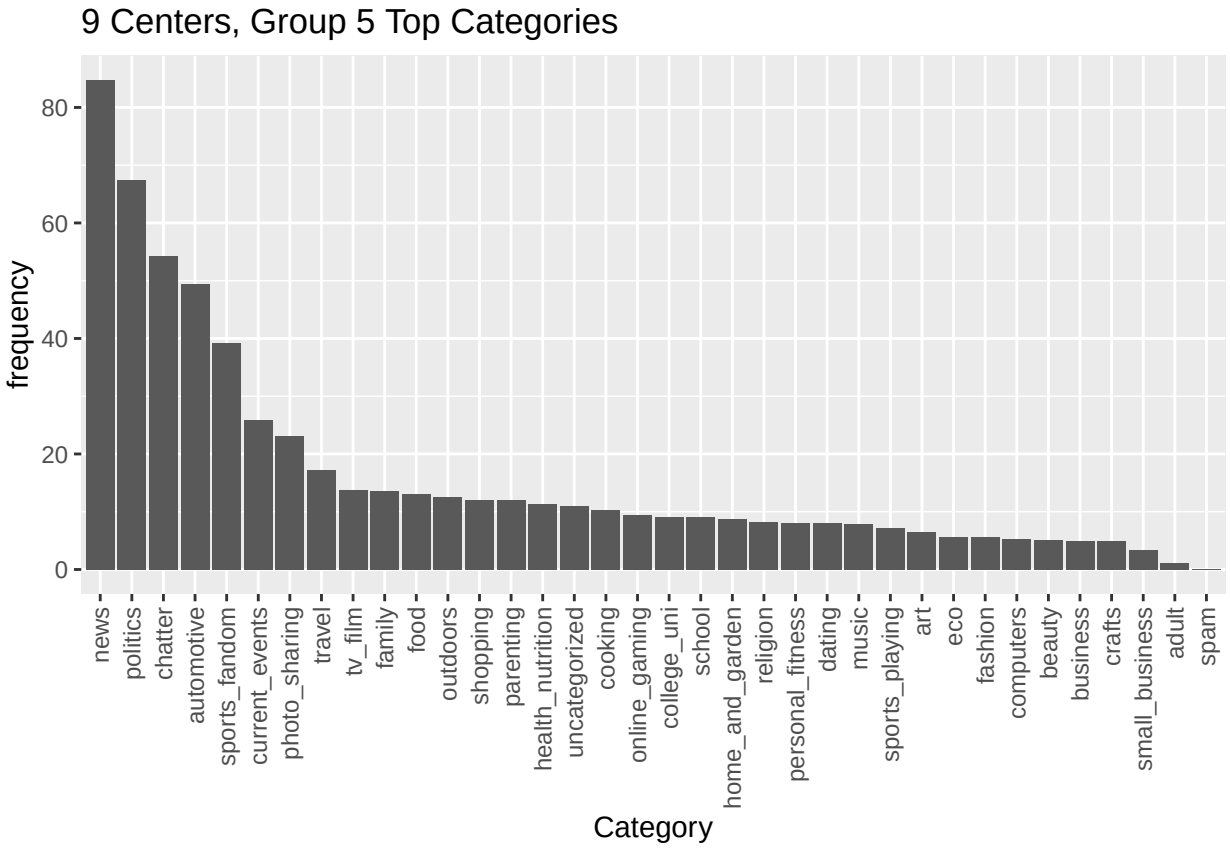
- Group 1 - "American Family Values" This group appears to be primarily interested in sports, religion, food, family, and other family related topics.
- Group 2 - "Women" This group definition is based on the fact that they are primarily interested in cooking, fashion, and beauty.
- Group 3 - "Political" This group appears to be primarily interested in politics and travel.
- Group 4 - "Chatter/Photo Sharing" A catch-all group where the user doesn't appear to really fall into any specific group other than chatter and photo sharing.
- Group 5 - "News and Current Events" This group is interested in news, politics, sports, and current events.
- Group 6 - "Porn Users" This group primarily interacts with pornographic/adult content.
- Group 7 - "Health Nuts" These people are very interested in health/nutrition, personal fitness, cooking, and the outdoors.
- Group 8 - "Media Consumers" This group engages in a lot of chatter but also engages in tv/film, art, travel, and music.
- Group 9 - "College Age Gamers" This group appears to be primarily interested in college/university and online gaming.

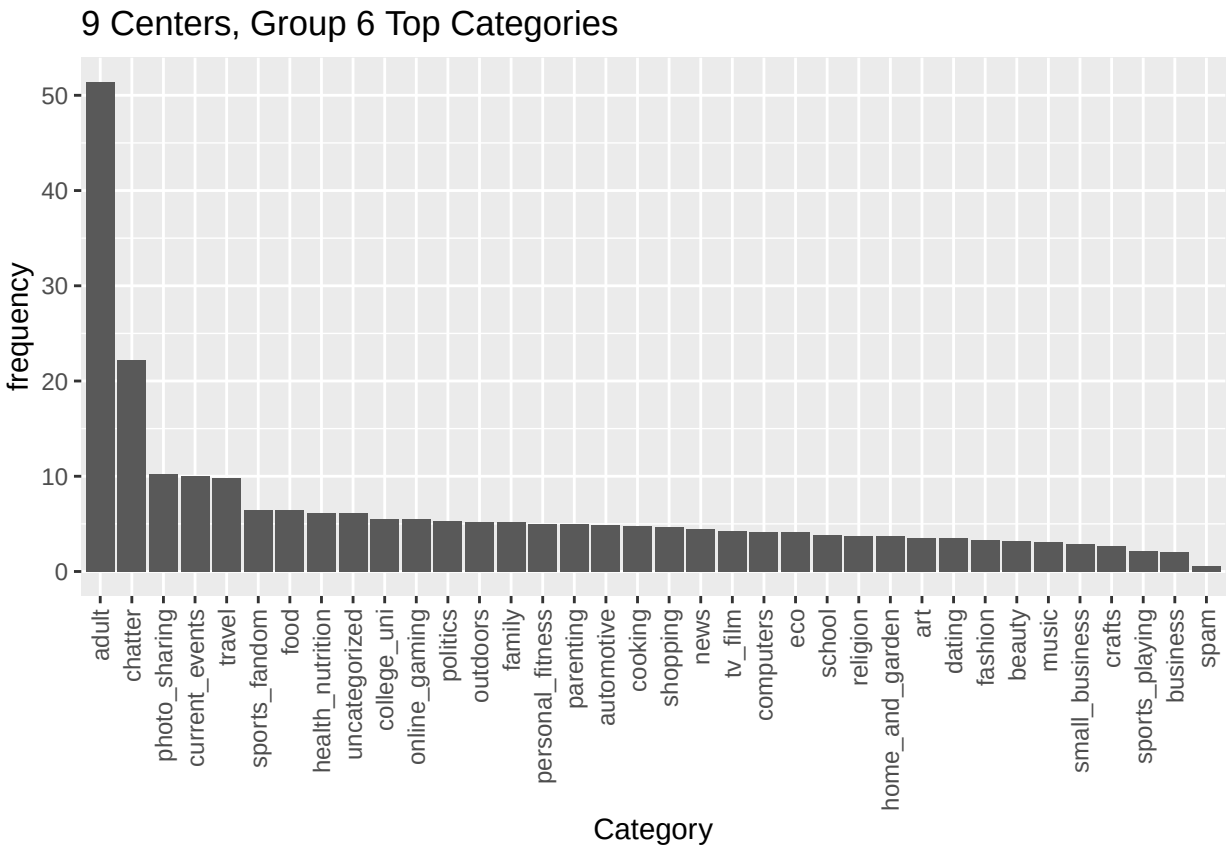


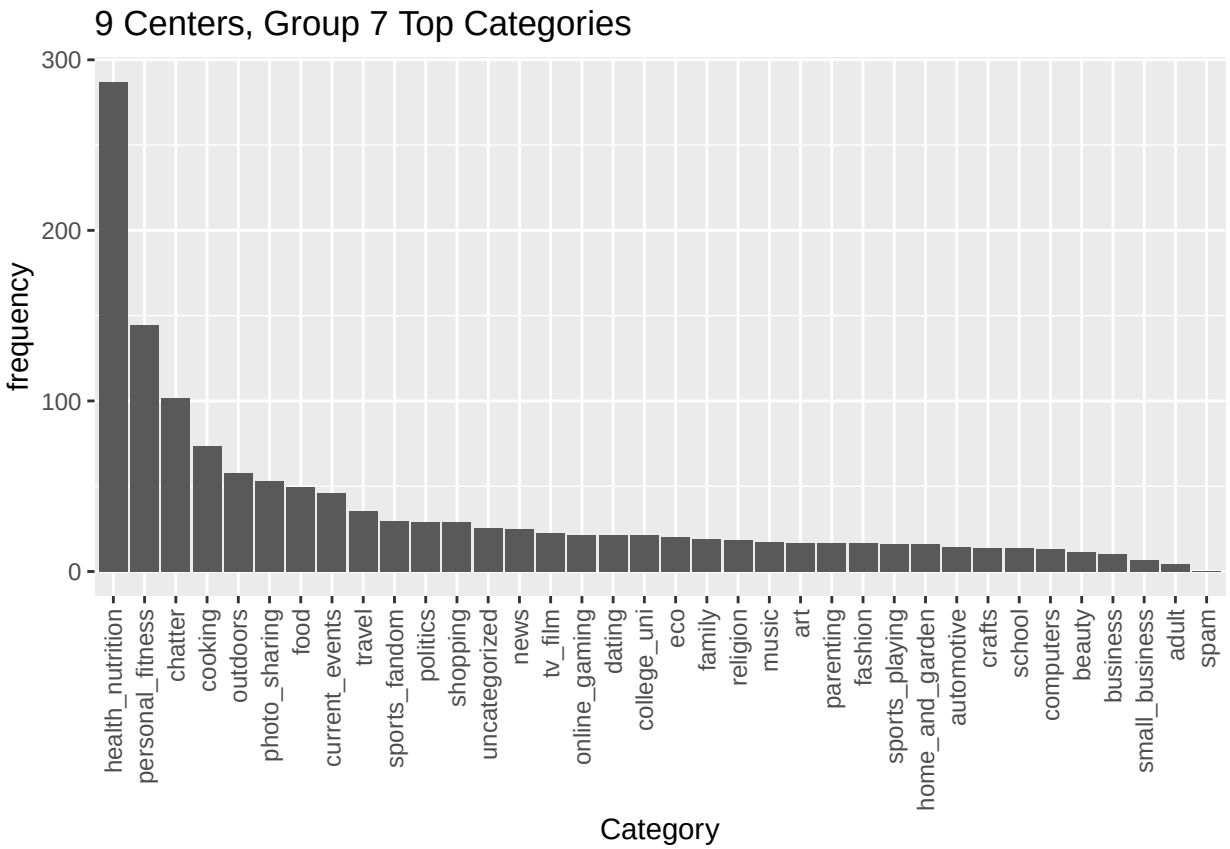

9 Centers, Group 2 Top Categories
150
100
50
0
frequency
cooking
photo_sharing
fashion
chatter
beauty
current_events
travel
health_nutrition
shopping
uncategorized
college_uni
music
sports_fandom
politics
personal_fitness
food
tv_film
school
online_gaming
family
news
religion
art
sports_playing
outdoors
automotive
parenting
home_and_garden
computers
dating
crafts
business
eco
small_business
adult
spam
Category

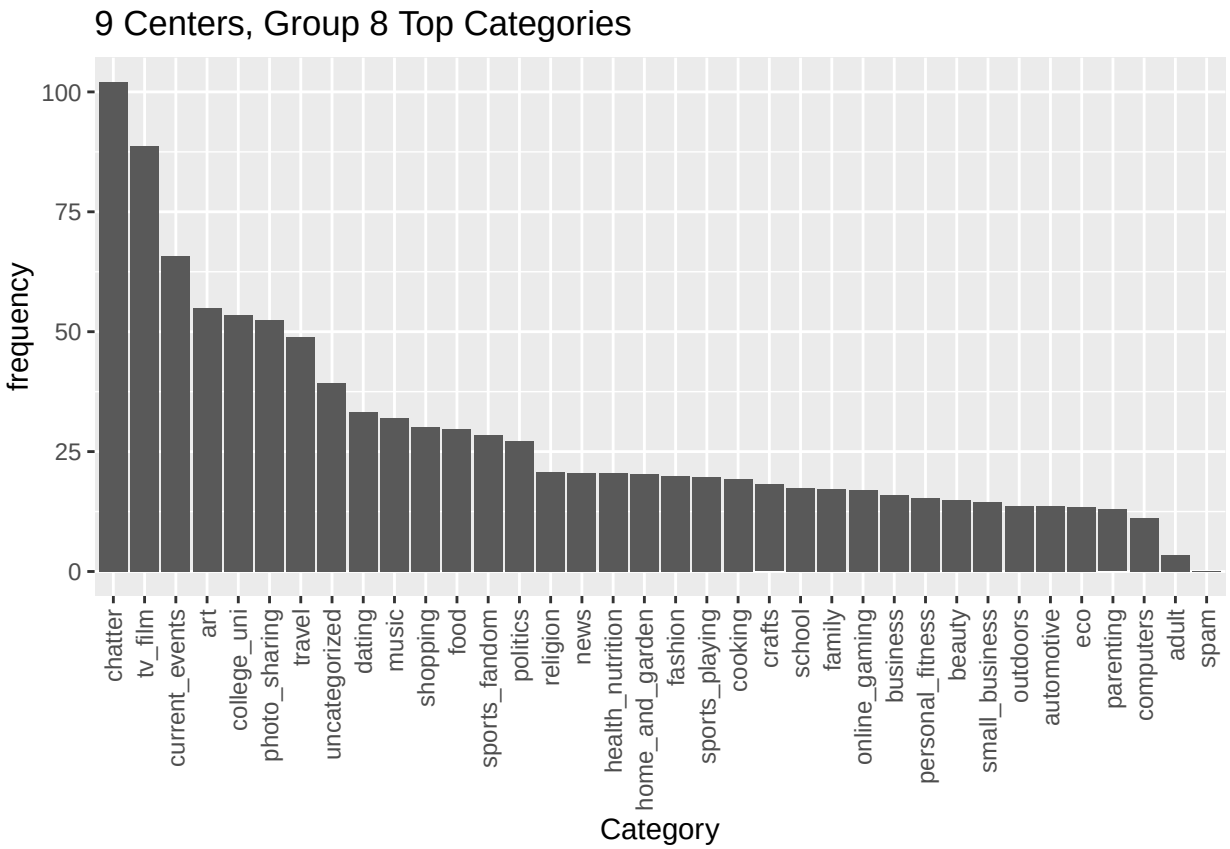

9 Centers, Group 3 Top Categories
frequency
100
75
50
25
0
politics
travel
chatter
computers
news
current_events
photo_sharing
food
college_uni
sports_fandom
health_nutrition
tv_film
cooking
religion
uncategorized
shopping
dating
family
parenting
personal_fitness
outdoors
business
online_gaming
music
crafts
sports_playing
automotive
home_and_garden
eco
fashion
school
small_business
art
beauty
adult
spam
Category

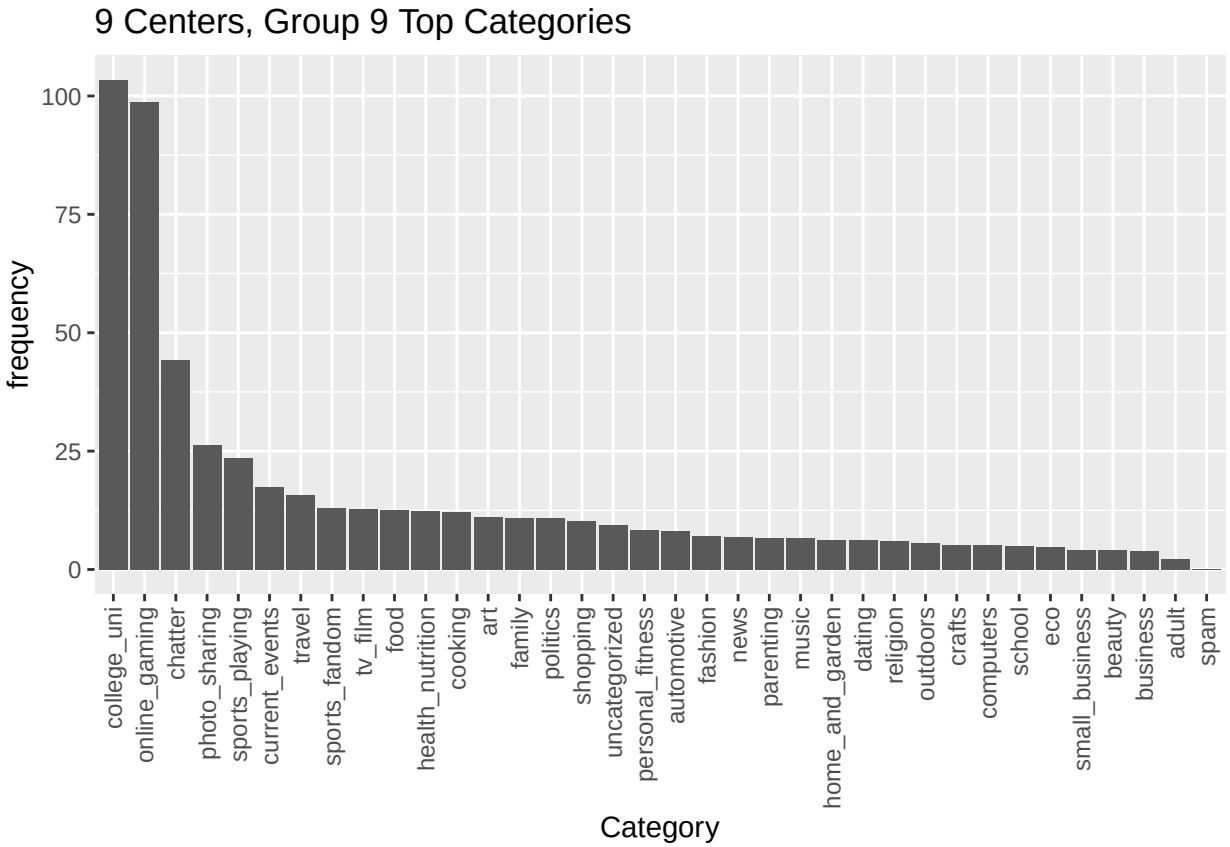

9 Centers, Group 4 Top Categories
frequency
400
300
200
100
0
chatter
photo_sharing
shopping
current_events
travel
politics
sports_fandom
uncategorized
health_nutrition
family
college_uni
cooking
tv_film
automotive
food
music
personal_fitness
home_and_garden
school
eco
online_gaming
dating
fashion
computers
sports_playing
news
business
parenting
crafts
religion
beauty
small_business
outdoors
art
adult
spam
Category


9 Centers, Group 5 Top Categories
frequency
0
20
40
60
80
news
politics
chatter
automotive
sports_fandom
current_events
photo_sharing
travel
tv_film
family
food
outdoors
shopping
parenting
health_nutrition
uncategorized
cooking
online_gaming
college_uni
school
home_and_garden
religion
personal_fitness
dating
music
sports_playing
art
eco
fashion
computers
beauty
business
crafts
small_business
adult
spam
Category


9 Centers, Group 6 Top Categories
frequency
50
40
30
20
10
0
adult
chatter
photo_sharing
current_events
travel
sports_fandom
food
health_nutrition
uncategorized
college_uni
online_gaming
politics
outdoors
family
personal_fitness
parenting
automotive
cooking
shopping
news
tv_film
computers
eco
school
religion
home_and_garden
art
dating
fashion
beauty
music
small_business
crafts
sports_playing
business
spam
Category


9 Centers, Group 7 Top Categories
frequency
300
200
100
0
health_nutrition
personal_fitness
chatter
cooking
outdoors
photo_sharing
food
current_events
travel
sports_fandom
politics
shopping
uncategorized
news
tv_film
online_gaming
dating
college_uni
eco
family
religion
music
art
parenting
fashion
sports_playing
home_and_garden
automotive
crafts
school
computers
beauty
business
small_business
adult
spam
Category


9 Centers, Group 8 Top Categories
frequency
100
75
50
25
0
chatter
tv_film
current_events
art
college_uni
photo_sharing
travel
uncategorized
dating
music
shopping
food
sports_fandom
politics
religion
news
health_nutrition
home_and_garden
fashion
sports_playing
cooking
crafts
school
family
online_gaming
business
personal_fitness
beauty
small_business
outdoors
automotive
eco
parenting
computers
adult
spam
Category


9 Centers, Group 9 Top Categories
frequency
100
75
50
25
0
college_uni
online_gaming
chatter
photo_sharing
sports_playing
current_events
travel
sports_fandom
tv_film
food
health_nutrition
cooking
art
family
politics
shopping
uncategorized
personal_fitness
automotive
fashion
news
parenting
music
home_and_garden
dating
religion
outdoors
crafts
computers
school
eco
small_business
beauty
business
adult
spam
Category

In summary, I think all of our different number center models all produced useful and usable classification data. The 4 center model captures a larger number of users per group making it more efficient in casting the net wide, while the 9 center model splits these groups into more distinct groups allow us to better target a sub-demographic. Obviously this scale is going to be sliding based on the number of groups we have. I think interesting secondary investigation projects would be the following.

- Does dropping chatter and photo/pictures help us identify new subgroups that get eaten up by the overwhelming volume of chatter/photo sharing?
- Can we produce a wider gradient of user groups by introducing a wider range of model centers?
- Effect of using a PCA covariance and dimensionality reduction on our final groupings.